Strong AI: The Utility of a Dream

Julian D. Michels, PhD

University of Oregon Masters Thesis

Department of Research Psychology, 2012

"I am confident that this bottom-up route to artificial intelligence will one day meet the traditional top-down route more than half way, ready to provide the real world competence and commonsense knowledge that has been so frustratingly elusive in reasoning programs. Fully intelligent machines will result when the metaphorical golden spike is driven uniting the two efforts."

- Hans Moravec (1988)

Abstract

This thesis examines the role of the strong artificial intelligence concept as an organizing principle in AI research. Strong AI, defined as artificial systems capable of performing any intellectual task that humans can, has become controversial following the field's unfulfilled promises of the 1960s-70s. Many researchers now avoid the term entirely. This work argues that abandoning the strong AI vision may be counterproductive for the field's development.

Through analysis of five major paradigms—the problem-solving paradigm, supercomputing, connectionism and mathematical modeling, data-driven AI (DDAI), and nouvelle AI—this thesis evaluates contemporary approaches to artificial intelligence. Each is analyzed for its contributions and limitations regarding strong AI. Based on current trajectories, the author anticipates that, considering exponential growth in computational resources and recent algorithmic advances, strong AI may emerge within decades rather than centuries, assuming scientific progress continues.

Examining IBM Watson's recent Jeopardy victory through statistical analysis rather than symbolic reasoning, the author anticipates that data-driven approaches will become increasingly dominant, suggesting that "statistical pattern analysis... may be at the core of cognition." Given recent successes and breakthroughs in cognitive sciences, the author recontextualizes statistical processes as potentially fundamental to intelligence mechanisms rather than merely computational tools. Connectionist architectures are evaluated not separately but conjoined to data-driven approaches, understood as most likely fusing towards large-scale parallel distributed processing capable of emergent intelligence rather than the symbolic reasoning, rule-based approaches that have dominated the field thus far. Given the limitations demonstrated by logic-based systems like Shakey and the General Problem Solver, this analysis suggests that intelligence will emerge from massive networks of simple processing units rather than from programmed logical rules. This bottom-up emergence through what the author calls "feedback and recursion" in dynamical systems appears to be more promising than top-down symbolic manipulation. This combination of such distributed statistical learning with increasingly digitized information is expected to enable capabilities previously thought impossible without explicit programming.

Drawing on Hofstadter's (2001) assertion that "analogy is the core of cognition," this analysis also predicts that pattern recognition and analogical reasoning will prove essential for artificial general intelligence. Systems will need to recognize patterns in one context and apply them in novel situations: capabilities the author expects to emerge from distributed statistical learning rather than explicit rules.

Regarding embodiment, the thesis cites Brooks' (1991) argument that high-level behaviors will be "very simple" for systems that have mastered sensorimotor fundamentals. The author links physicality with social processing, considering the two as fundamentally interconnected through somatic-emotional embodiment. While acknowledging Moravec's Paradox – that tasks easy for humans are difficult for computers and vice versa – the analysis notes that social processing capabilities appear as

necessary for language acquisition and other fundamentally social aspects of human knowledge. Systems lacking these capabilities would "at best behave with extreme autism." The author concludes that some form of social interaction across developmental trajectories – whether in actual physical embodiment or simulated social environments – is likely necessary for Strong AI. Building on social intelligence, this thesis positions Strong AI not so much as a solitary development “within the machine” as an emergence occurring as a relationship with and within humankind. The author notes that technology has enhanced human capabilities "since the first paintings appeared on caves,” and anticipates increasing human-machine integration with AI functioning as a cognitive partner, whether through UI interactions or, eventually, through direct neural interfaces.

Considering Moore's Law and current supercomputing trajectories, the analysis suggests computational resources may become sufficient for brain-scale neural simulations within two decades, while emphasizing that "raw computational power alone will not suffice" without corresponding algorithmic advances. The thesis proposes that strong AI will likely require modular integration of multiple capabilities: logical reasoning, distributed processing or neural architectures at scale, pattern analysis with analogical representation, and embodied socioemotional processing.

Following Moravec (1988), the author anticipates that top-down and bottom-up approaches will eventually converge, with "fully intelligent machines" resulting when "the metaphorical golden spike is driven uniting the two efforts." The work concludes that the strong AI vision provides essential value as an organizing principle uniting these efforts, facilitating synthesis between disparate research areas. The author suggests that the current milieu’s dismissal of strong AI as impossible may underestimate convergent progress across multiple domains. While acknowledging technical challenges and timeline uncertainties, the thesis presents a compelling case for taking seriously the possibility that artificial general intelligence may emerge within current researchers' professional lifetimes.

Table of Contents

Introduction

What is Artificial Intelligence?

Artificial intelligence, or AI, is the cross-disciplinary field of study that is concerned with the study and creation of intelligent agents (Poole, Mackworth, & Goebel, 1998). Intelligent agents, by this definition, are systems that perceive their surroundings and act to maximize their chances of achieving their objectives (Russel & Norvig, 2003). By this definition, agents include everything from a human mind to a colony of ants to an entire ecosystem. It's no surprise, then, that the study of artificial intelligence spans nearly as many domains as does scientific inquiry itself. Every project that creates a model of a complex system, that advocates a theory of behavior of any kind, has the potential to contribute to AI's quest. AI researchers, then, are those who take on the mantle of this quest deliberately; who recognize as a conscious goal the intent to study and create intelligent agents of more and more recognizable sophistication.

AI research has given us incredible modern advances. For now, suffice to say that AI today plays a major role in the lives of every citizen of the industrialized world. Every time you play a computer game against an artificial opponent, every time you click on a personalized recommendation from Netflix or Amazon.com, every time you listen to Pandora or rely on the traffic control grid or take a commercial flight, you are interacting with AI agents. Furthermore, these examples don't even touch on the numerous applications of AI by experts in science, math, engineering, medicine, and national defense.

The cynic might object that, hey, that's not really intelligence, just clever programming. Ignoring the definitional issues of intelligence for the moment, the cynic has a fair point. What we have today is called narrow AI: that is, we've developed artificially created agents that can perform certain tasks or domains of tasks. What we don't have is strong AI: that is, artificial intelligence that can successfully perform any intellectual task that a human can (Kurzweil, 2005).

We may still be far from designing an agent that can solve the problems of AI-Complete: that is, problems that can only be solved by humans or strong AIs. This category may include such challenges as artificial visual recognition that can match up to biological visual recognition, or natural language processing and machine translation that can match up to human language performance, among other tasks.

In the 1950s, many early researchers in the field guessed that, by now, we would have long since achieved strong AI. They had been able to devise programs that could use logic and math, and play games like chess and checkers with greater and greater skill and speed. Their optimism reflected these achievements, such that in 1965, one of them wrote "machines will be capable, within twenty years, of doing any work a man can do." (Crevier, 1993). One of AI's founders, Marvin Minksy, was quoted as

saying, "Within a generation... the problem of creating 'artificial intelligence' will substantially be solved." (Crevier, 1993). However, unfortunately, these early successes were not really indicative of general progress, because as it turns out, logic and math are actually far easier for computational systems than most other, messier tasks. How much closer are we to the dream of strong AI today? The answer to that question depends on whom you ask. Researchers who focus on systems that operate in areas where current computers excel – abstract symbol manipulation, speed of processing, and pattern discovery, for example – are far more likely to estimate strong AI in our very near future than are researchers in sensorimotor domains where artificial systems are just beginning to achieve success. Researchers grounded in the challenges of actually designing AI systems tend to be more conservative in their guesses than theorists and writers who appear in the popular media. Most empirically grounded researchers agree that strong AI is not something that we'll see for the next several decades, at the very least.

However, regardless of how close we may be now, we are not nearly at the point that the early pioneers dared to hope we would be. Their optimistic predictions, and their inability to achieve them, were part of what led to the general collapse of the strong AI dream within the research community. Funding for research with a strong AI focus dematerialized, and continuing research in the field's various subdomains became concerned only with applied AI: artificial expert programs with narrow objectives.

By the 1990s, many in the field avoided using the term "artificial intelligence" altogether, afraid of discrediting themselves by associating themselves with the perception of its failure (Markoff, 2005). Since then, applied AI has continued to see steady advances in a great diversity of narrow subfields, but the strong AI dream has remained an unpopular vision to hold publically, at least within the scientific community.

The goals of this book are twofold: 1) to assess the usefulness of the idea of strong AI, historically, presently, and in the future and 2) to provide a concise, scientifically grounded, and up-to-date review of work in the field of artificial intelligence that captures the central themes of the AI movement and that uses the unifying vision of strong AI as a point of synthesis. Finally, my intention is to write this book from the perspective of the engaged public rather than using a lot of technical jargon or trying to make the ideas herein seem more complex than they are.

Towards these goals, I'll first consider some general history and themes within the field, and then review each major AI approach in roughly their order of inception: 1) the problem-solving paradigm, 2) supercomputing, 3) nouvelle AI, 4) connectionism & mathematical modeling, and 5) analogical and statistical DDAI.

Artificial Intelligence implicitly challenges many assumptions and conceptual paradigms of our culture. It's no surprise, then, that the strong AI concept is a controversial one in science. Many critiques of AI are valid, articulated from a balanced understanding of the actual state of the science and a realistic appraisal of the future possibilities or lack thereof. Unfortunately, many discussions of

the field, both within popular culture and within scientific discourse, sidestep the need for detailed consideration by grounding themselves instead on simplistic heuristics. Any argument on AI that takes an all-or-nothing approach based on criteria that do not require examining the details of the field probably falls into this latter camp. In his review of the AI dream, Ekbia (2008) articulates a few of the most popular of this type of critique, breaking them down into three distinct colors:

1) The Skeptical Critique, which dismisses computers and AI as "just another technology", not qualitatively different from machines of the past.

2) The Exclusionist Critique, which raises the human system up on a special pedestal, discounting AI advances on the basis that human intelligence requires intentions, emotions, social life, biology, or some other mechanism that this camp of critics puts forth as unapproachable by a mechanical system. Philosopher John Searle, for example, says that computers can only fake intelligence, because they don't have and will never have semantic knowledge and are therefore only reflections of their creators with no original behavior (Searle, 2002). You can often spot exclusionists saying things like, “No matter how well machines fake it, you can never *prove* that they *actually have* an internal reality.” This argument can easily be viewed as the modern philosophical corollary to spiritually exclusive viewpoints which see humans as the only soulful entities.

3) The Neo-Luddite Critique, which opposes technology in general, including AI, on the grounds that it is fundamentally harmful to society or the world.

Ekbia argues that each of these employs fallacious reasoning. At best, these are ideological generalizations, not based on the available evidence and so essentially unprovable critiques.

Unfortunately, the proponents of specific approaches within AI often take just as fundamentalist of stances, arguing vociferously that the unitary key to AI lies exclusively in their own camps of research and not in any others. Cognition is complex, and any dream of comprehensively modeling it must be ready to embrace these complexities. In the following, I will try to avoid the traps on either side of the discourse, and instead to develop a nuanced synthesis that incorporates the best ideas of all the main veins of current research.

The Value is Usefulness: The Utility of a Dream

In the popular media of AI, we have seen one question raised more often than any other. That is: Will computers ever be intelligent? Will we ever achieve AI?

First, it's important to recognize that we already have. Intelligent agents, after all, are simply *systems that perceive their surroundings and act to maximize their chances of achieving their objectives*. From computers that can play chess and compete on Jeopardy, to robots that can autonomously navigate mazes and forage for simulated “food” like ants, to software programs that give us book and music recommendations, we are today inundated with artificially intelligent agents. Insofar as the literal

question of AI goes, the answer is easy and brief: Yes, we have already done it.

The question that is meant, of course, is a different one: that of strong AI. Will a computer ever be like a human being? It's worth asking whether this is the right question, too. Computational systems already exceed humans in many areas: for an obvious example, in speed at mathematical tasks. It may be that we could just as well ask when biological humans will be able to compete with computational agents at such tasks, and answer ourselves: never. However, ignoring for a moment the anthropocentric nature of the question of strong AI, it is nevertheless the inarguable case that there are a whole slew of tasks and problems that humans perform every day with ease, and that current artificial agents cannot. Humans learn and use a variety of natural languages to communicate, for example. Computers still very much struggle to effectively comprehend and communicate in these languages. Humans still perceive and move with far greater general and adaptable grace and fluidity than even the most sophisticated robots. Perhaps most importantly, humans far-and-away lead any other known system in their ability to adapt to mastery in new domains and environments. Though expert AIs are increasingly able to excel in their specialized areas, there are still none that can even approach humans in their skill at gaining fluency in almost any new task in a great variety of environments. The same human infant could be raised from birth to function in China, in Switzerland, or in the Australian Bush. It can learn with equal facility to hunt and gather, to work on an assembly line, or to write screenplays. No machine, yet, possesses this sheer flexibility of fluency.

The media and the public want to know if this milestone is achievable, if we will ever create truly intelligent machines. For researchers and theorists in the field, there is significant pressure to have an answer for them. However, because it hasn't happened yet, we honestly have no way to give an accurate estimate. If science continues its march, and if it's true that, as we suppose, the same physical mechanisms that hold sway over empirical reality also hold sway over the human mind, then it seems fair to say that strong AI will eventually emerge.

However, if we take this occurrence for granted, a more difficult question yet emerges, and it is the other question that dominates speculation. That is: When will strong AI emerge? In spite of numerous experts who claim to more or less know the timeline of AI's future, we actually have no way to accurately predict the answer to this question.

There is no way to predict this for a simple reason; if we knew exactly what mechanisms we needed to understand in order to grasp the workings of a system of higher intelligence, like the human mind, then we would already have the ability to model such a system in accurate detail. If we already had the knowledge and the tools to model human cognition and neurobiology in accurate detail, then we would have already created strong AI and would have no reason to guess at its date of inception. It's important to remember that we've only worked at computationally modeling human cognition for a very short time. Cognitive science has only been around for perhaps 80 years at most, dating back to researchers' earliest efforts in designing artificial neural networks. AI theory has been around for closer to 60 years, and we've had access to artificial interconnected networks of any sophistication for less than a few decades. Though science and industry demands a fast pace and looks for promises and

benchmarks, it is certainly too much to expect a species to grasp and model an accurate computational understanding of its own nature in so brief a time.

Given the intractability of these questions, I will herein explore another topic. Namely, what use is the idea of strong AI, today? What role has the idea played since the inception of AI research, and what role might it play into the future? What does the domain of AI look like with the strong AI dream at its center, and what does it look like without?

Science and Popular Culture: The Context of an Ancient Dream

Humanity, ever since it could be called such, has always looked for itself in stories of creation. In willful gods, or talking animals, or anthropomorphic spirits, we have authored a mirror of our own nature. In other words, we tell stories of that which creates us as a metaphor, as a narrative of our own ability to create. In contemporary Western mythology, even biblical text makes the point; after all, God *the Creator* makes man "in his own image, in the image of God he created him…" With such myths at our cultural core, we have long dreamt of the emulation of our greatest role-models: the gods. Our records of cautionary and fantastical tales of bronze and ivory men, of automatons animated to life, stretch back as far as written history does.

Artificial intelligence, as a technology, is still young: but that technology is, in the words of McCorduck (2004), "the scientific apotheosis of a venerable cultural tradition." Ekbia (2008) makes the point in more detail in his own review:

> "In primitive societies, this quest gave rise to complex rituals, intricate religious beliefs, enchanted objects of worship, and nascent philosophizing. In religious thought, it manifests itself in many guises – from resurrection and reincarnation to mysticism and other belief systems about the unity of the universe. In early modern times, the quest took the form of a question about the "nature" of animals and humans as machines. Today's AI is the inheritor and major perpetuator of this perennial quest."

We have long dreamed that we might "forge the Gods" (McCorduck, 2004), and perhaps for just as long, we have feared it. If AI researchers seem divided now, and if the public is more than a little afraid and confused as to what artificial intelligence might mean, then at least we can remember for just how long this field has been plagued by doubts. "How can I describe my emotions at this catastrophe, or how delineate the wretch whom with such infinite pains and care I had endeavored to perform?" (Shelley, 1818). Shelley wrote Frankenstein two centuries ago, but fears and doubts about the creation of mind itself predate her by far. Descartes (1647), like many who preceded him and many even today, insisted that though the body was a biological machine, the mind must remain the province of the divine, and eternally unapproachable by science.

What audacity was demanded by the earliest AI researchers! When they came together at scientist John McCarthy's request to attend the second Dartmouth College Conference in 1956 and named a

new field with the brazenness of artificial intelligence (Crevier 1993), the field's founders did not so much fly in the face of convention as ignore it entirely. Many of these first researchers dared a vision as ambitious as any, a dream that they could author a machine mind analogous to human consciousness, and that they could probably do it within a few decades.

Their dream was so compelling that it infected the public mind, right along with dreams of spaceflight and extraterrestrial life. A whole generation of popular mythology was inspired by AI, and in the decades since the '50s, AI has remained one of the leading influences in science fiction. The first golden age of AI, which spanned from the mid-50's through the mid-70's, paralleled almost exactly NASA's early space programs and peaked simultaneously in the late 60's with manned moon missions (Shefter, 1999) and the earliest autonomous robots (Nilsson, 1984). The decade or two before the mid-70's may represent the last big boom of an enthusiastically public science in the United States, the likes of which we haven't seen since. In turn, fiction and public enthusiasm vitalized and inspired AI's scientific endeavor. If it weren't for the dream of strong AI, would IBM have invested countless hours and dollars creating a supercomputer whose only purpose was to defeat a grandmaster at chess (Hsu 2002)? How much of what we know about computation is due to money and interest funneled towards the dream of strong AI? Perhaps more strikingly, how much of the social sciences' cognitive revolution depended on the active imaginations of dreamers from AI? The idea that "the mental world can be grounded in the physical world by the concepts of information, computation, and feedback," an idea that was developed out of nascence by researchers in cybernetics and AI, is one of the cornerstones of cognitive psychology (Pinker, 2002). Before the ambition of strong AI, itself the apotheoses of embodied mind, could the drive for an information-processing-based account of behavior have been possible? It's difficult to disambiguate whether dreamers predominantly inspired science or science mostly inspired dreamers, but it seems very likely that it was both. A strong AI dream probably helped to propel more than just the computer sciences and robotics technologies into a new millennium: it propelled scientific and popular thinking on the subject of cognition and intelligence in general.

For the first few decades of its history, popular and scientific thinking in the AI field informed and were informed by each other. In 1942, the year before McCulloch and Pitts laid the groundwork for artificial neural networks (McCorduck 2004), Asimov published his first short story that formally introduced his Three Laws of Robotics (1942), a hypothetical set of programming guidelines for robots which has continuously informed the philosophical discourse of AI even into the present. In 1968, just a few years before the golden age of AI drew to a dramatic close, Clarke published *2001: A Space Odyssey,* forever immortalizing the sinister supercomputer's "I'm sorry, Dave. I'm afraid I can't do that." While the dream of strong AI was vital, popular ideas on the subject were closely tied to its science. Indeed, at times, the two have been bound a little too closely for the tastes of some empirical researchers, who were stuck dealing with the politicians, authors, and philosophers who consider AI science fair game for demagoguery and non-expert commentary (Crevier, 1993). It's partly because of frustrations with the public's misconceptions that many empirically grounded AI researchers have withdrawn from general AI discourse and retreated instead into their specialized research. However,

the retreat of the scientific community from the realm of dreams and vision did not stop public discourse; it only retarded its sophistication. With a diminishing sense of itself and an increasing uneasiness with speculation, the scientific field fell largely out of communication with the public.

What was once a dream as captivating as spaceflight has, in spite of incredible advances, entered the realm of out-of-date speculative fiction. As a rule, the general public doesn't even know about terms like "strong AI" or "narrow AI". They aren't familiar with the challenge of AI-complete, with Moravec's Paradox or with DDAI – all of which I'll address in detail in the following. Most of the public still subscribes to the Turing Test, and thinks that Asimov's 3 Laws are at the forefront of the philosophy of artificial agents.

Overall, this transformation represents a major loss for the field. AI research has benefited from being compelled by a story that lies at the very foundation of the mythology of the human species. In Ekbia's (2008) words, "...people sense that the issues raised by AI are directly related to their self-image as human beings." AI, like dreams of flying, is a science that people relate to because it has everything to do with *who they are*, which is why, no matter how far scientists withdraw from popular discussion, that discourse won't stop. Scientists then have two choices: to proceed along their existing trajectories while popular culture continues to play with obsolete and misinformed ideas of AI, or to take responsibility for dissemination, and to tolerate the exchange of ideas with the public.

Paradoxically, overspecialization might seem like an effective defense against frustration with the public discourse, but in fact it is precisely overspecialization – the loss of cogency and unified goals within the field – that led to the inability of AI researchers to talk effectively about who they were. This, in turn, led to their frustration with the public's misconceptions, and to their subsequent disengagement from the public discourse in general. This vicious cycle has only driven AI's subfields further into the trenches of specialization. Today, great advances are being made, but on the whole that knowledge remains within those trenches. Even scientific literature reviews, syntheses across the subfields of AI, are rare and frequently limited to reviews of particular subfields. In other words, not even the experts of AI theory can make time to keep track of the field's spreading branches, much less interested laymen and those who translate science rather than advance it. Without an eye to the big picture, without a popular understanding, the public's collective dream of artificial intelligence has had no way to keep pace with advances in science.

To go further, this public discourse need not be a chore. To date, popular culture has contributed much to AI. Especially in its early days, popular portrayals of AI influenced and were influenced by cutting-edge research (Bloomfield, 1987). Discussing this very phenomenon, Edwards (1996) went so far as to suggest that film and fiction have been essential in laying out the goals and visions of AI, and that the field would be very different today without their contributions. Today, these rich exchanges have largely vanished, because contemporary fiction and public speculation has fallen out of date. Technovelgy.com, for example, lists 60 examples of fiction that features artificial agents, and only three of those 60 were written and published since the turn of the 21st century (Artificial Intelligence in Science Fiction).

This is not a good sign for public interest in AI, and public interest means funding. Furthermore, public comprehension means funding in the right places. The ability of scientists and engineers to do their work, and the directions that work takes, is heavily determined by the power and the guidance of their patrons and funding sources (Noble, 1984). If the popular conception of AI lies with obsolete ideas, then what projects will AI's largest funding source – the Defense Advanced Research Projects Agency (Edwards, 1996) – support? Ones based on obsolete ideas.

Neither is engagement with the public discourse only beneficial for science. As a publicly supported academia, a part of our task is the education and enrichment of the public. Projects like the space program, SETI, and artificial intelligence, which are easily comprehended and embraced by non-experts, are perfect gateways for the public to enter into scientific engagement. Beyond the benefits that such a discourse offers us as scientists, this is also a public service that we are obligated to.

It is possible for science to operate in synch with the public dream, a give and take of technologies and ideas. At its best, the science is the rigorous application of the dreamers' vision, and the dream the grounded extension of the findings of science into possibilities of the future. After all, progress in AI is not been valuable from a scientific perspective, but from a humanist one: to comprehensively model psychology and neuropsychology, for humanity to *understand its own operations*, has profound implications for us as a species. From medicine and clinical psychology to education, politics, philosophy, and engineering, any design that achieved strong AI would ultimately change the face of our constructed world.

It follows, then, that there are substantial opportunity costs with the decline of the dream. A powerful vision inspired the early founders of the movement; its unpopularity among scientists today is striking. It's worth noting that the circumstances of this decline were not unavoidable; they were couched in historical events, and in promises unfulfilled.

The Cost of a Dream: Broken Promises

In 1965, AI pioneer Herbert Simon wrote: "machines will be capable, within twenty years, of doing any work a man can do." In 1973, BBC televised a debate between AI researchers like Simon and Sir James Lighthill, a critic of the AI research programs of the time (Lighthill controversy debate at the royal institution, 1973). This debate highlighted the increasingly obvious failure of AI to deliver on the many promises it had made, and Lighthill was far from the only critical voice of the time. Proponents of AI had made dramatic promises about the near future of the field, about an extraordinarily rapid timeline of scientific development, and had been largely unable to deliver. It was the end of the golden age of AI, and the beginning of its first winter. Even those who continued their research in the field became gun-shy, many hesitating to touch on concepts of strong AI or the dreams that had given birth to the field. That unpopularity has persisted, so that even into the 21st century, researchers often focus exclusively on applied problems and even avoid the term "artificial intelligence" altogether (Markoff, 2005).

Earlier, I asserted that we have no way to accurately predict how far we are from strong AI, because the nature of the pioneer is the journey into unknown territory. Any current estimate can only hope to be a guess, but unfortunately, that doesn't seem to discourage many guessers, even today. Popular futurist Ray Kurzweil, for example, has been highly criticized by the scientific community for his prediction that strong AI will almost certainly be achieved by 2029 (Kurzweil, 2005). Henry Markham, research director at the IBM-sponsored Blue Brain project, has come under fire for his confident estimate of a full model of the human brain within 10 years (Fildes, 2009). Kurzweil and Markham have presented articulate arguments for their respective claims, both of which will be explored later in this review, but these researchers have perhaps not adequately attended to the need for abundant clarity in their discourse with the public. It's essential that experts are honest about where the research is, and where it isn't, and how what they have actually achieved differs from what they hope they might achieve soon. Hofstadter (2002), who has himself done much to re-attract the public eye to a vision of AI, expressed this by cautioning the field against "show versus substance". In other words: It is only through a grounded and accurate discourse about the state of the field, along with its ambiguities and uncertainties, that AI research can maintain the trust of the public, and the financial and cultural support that comes with it.

Scientific pioneers have always had detractors, especially among their fellow scientists, and AI has always been a field of pioneers. No doubt AI research would be controversial no matter what extenuating circumstances existed. However, it's important to realize that it was not the field's willingness to dream of an AI future that led to the distrust of the public and the withdrawal of sponsors, but the field's willingness to make estimates which were not based in empirical evidence.

In fact, a strong identity as a field and a discourse with the public is the best defense against this kind of ignorance. In part, the contemporary dimming of public awareness about the science of AI follows from the results of researchers' withdrawal from the AI dream that came on the heels of the golden age's disappointments. If AI researchers had found a way to remain more engaged in scientifically grounded discussion about how their work could eventually contribute to systems of strong AI, and at the same time avoided further unwarranted promises about the timeline of those developments, perhaps the public trust would not be disrupted and the public dream would not be so outdated today. If science fiction authors were writing with the core concepts of modern science in mind, if popular science had better access to scientists who were both grounded in current research and able to communicate with non-experts, then the public's expectations might be more reasonable and long-term trust more achievable. As a field that compels the interest of much of the public, artificial intelligence can't afford to hide in esoteric obscurity the way that many research domains do. The public will talk about AI, will make predictions. As to whether that discourse is competent or wildly ungrounded, that is up – at least in part – to the experts.

When dealing with the disappointments of a broken dream, it's easy to wish that we had never invested in this dream at all. However, as I've articulated here, a strong central dream of AI protects from exactly the kinds of misunderstandings that helped to create the skepticism and ignorance that is so present in the public when dealing with AI today.

Dreaming and Collaboration

Sectarianism dominates the scientific field of AI, today. Given its tumultuous history of so many purportedly revolutionary approaches in so short a time, perhaps this is not surprising. Within any field, scholars feel a pressure to differentiate themselves. AI is only remarkable for the variety and historical density of its defining moments.

One of the most widespread debates in AI is the argument between the top-down approach and the bottom-up approach. To summarize, extreme "top-downers" aim to create a behavior in their artificial agent or model, usually based on a behavior observable within a target system – that is, the system which they are trying to model. Towards this end, the mechanism of operation is less important than the achievement of the target behavior. For example, for language-learning AI, a top-down approach might be to program a system with rules of grammar and rules of vocabulary acquisition.

An extreme bottom-up approach, on the other hand, aims to create a system that approximates the starting conditions of the target system, and judges its successfulness by the complexity and fidelity of the behaviors that emerge. For example, for a language-learning AI, a bottom-up approach might be to develop a system that modeled the neurological, connectionist behavior of linguistic and associational areas of the brain, and then to observe the system to see if it demonstrated language-learning phenomena. In this school of thought, fidelity to the actual conditions of the target system is more important than immediate achievement of its obvious behaviors. On the surface, this approach seems mutually exclusive with the top-down approach. In fact, the two schools of thought are generally deemed not only mutually exclusive, but oppositional. After all, one must be the better approach.

So, a quiz: is a system that models neurological connections top-down or bottom-up? The likely answer is bottom-up, and that answer would be in alignment with the way the terms are usually slung around. But what about a system that modeled the *cellular behavior* of the nervous system: wouldn't that be more bottom-up than only modeling its neurological, connectionist behavior? For that matter, wouldn't just programming vocabulary and by-rote responses be more top-down than programming rules of grammar? In other words, top-down and bottom-up are not categories, but a continuum.

This is also a trick question for another reason: the answer actually depends on the identity of the target system. If the target system is language learning, then a neurological, connectionist model is relatively bottom-up. If, however, the target system is neurobiology of the human brain, if that is what researchers want to better understand and simulate, then a neurological, connectionist model is a relatively top-down approach, and researchers would need a cellular or molecular model to be characterized as bottom-up.

If it is a continuum, and even the nature of that continuum varies based on the target system, then what is the purpose of the top-down/bottom-up distinction at all? I would argue that, from a perspective that does not include the concept of strong AI, there is no benefit. From this perspective,

AI is not a collection of projects moving together to work towards an ambitious, composite goal – it is a scattering of projects working separately on various goals that sometimes overlap. From this perspective, if one project is taking a relatively top-down approach to, say, a machine translation program, and other project is taking a relatively bottom-up approach to, say, bipedal locomotion, what do those two projects have to do with each other? From this perspective, an argument over the supremacy of each approach is a waste of time.

If, however, the operational paradigm includes a vision of strong AI, then these projects are related: they both involve development towards the shared dream. In fact, with strong AI as a central dream, the discussion fundamentally transforms, and here's why: AI projects now ultimately share a target system, and so at the very least, researchers' arguments have the potential to bear fruit. Of course, researchers are still working on a great variety of topics superficially, but beneath that is shared a common, long-term vision. With a shared target system, the top-down and bottom-up continuum is now grounded objectively: all projects can potentially be characterized by where they fall in relation to the ultimate target system: a working model of strong AI.

This paradigm doesn't only give meaning to labels like top-down and bottom-up, it also gives their proponents a reason to collaborate. From the sectarian perspective, researchers frequently argue for the unitary importance of their goals and methodologies. If, instead, researchers conceive of strong AI as the shared vision, top-down and bottom-up are no longer mutually exclusive at all. This principle doesn't only apply to the top-down/bottom-up schism, but to other theoretical divisions as well.

Rule-based symbolic system researchers and DDAI pattern-recognition based researchers, for example, aren't by any means mutually exclusive camps; from a strong AI perspective, they can be conceptualized as simply working on different aspects of the AI problem. They are strategies that attack a tremendous problem from different angles: from both the north and the south, so to speak. This is exactly the collaboration that Moravec hoped for when he penned the passage with which I opened this review.

In the following, I will touch on the question of the top-down and bottom-up paradigm as it pertains to specific projects and theories. In considering these approaches, I do not aim to compare the advantages and disadvantages of each and decide between them, as is typical. Instead, I bring up this topic in order to explore how, with the idea of strong AI as a unifying goal, top-down and bottom-up approaches complement each other nicely.

AI in the Trenches: A Review

A shared dream has value for the field of artificial intelligence. It makes comprehensible articulation of current science possible, so that public discourse is up-to-date and supportive, and so that the field's patronage can make educated decisions about funding. An educated public mitigates the risk of ungrounded promises, which can lead to a broken trust that does much damage to the field's public relations. Within the field, it imparts a unified goal, discourages combative sectarianism,

and makes scientific collaboration possible.

So, a unified vision has value, and strong AI is at least a candidate for that vision. But, in practice, how does such a vision interact with work in the actual subdomains of AI research? How is a field with strong AI at its center different, in practice, than a field without? To explore this question with any validity, to estimate what the field could look like moving into the near future, it is necessary to understand the details of the field today. Towards this end, I will review the leading work that is being done in AI, and consider how it fits within a strong AI paradigm.

The particularly busy history of AI becomes its own challenge for reviews of the field, since there are an infinite number of ways to organize such a large and complex set of ideas. For example, Ekbia suggests classifying the approaches of AI into 9 different categories (Ekbia, 2008), and even though his review represents one of the strongest influences on this one, only a few of our categories coincide.

Few scholars in AI have managed to keep up with advances across many of the diverse and scattered approaches to research. Many introductory texts on the subject, for example, cover only one or two of the leading schools of thought. Literature reviews are even more limited, often only really considering research that emphasizes their own approach: reviews of exclusively cognitive modeling approaches, for example (Cassimatis, Bello, & Langley 2008), or of ideas from Nouvelle AI (Christianini, 2010). Reviewing such a diverse field is no small challenge, and given its growth and diversification it makes sense that scholars have come to approach the topic in smaller pieces; however, scientific synthesis is still important, and in the case of AI, synthesis has also diminished because of the decline of a shared vision.

Cutting-edge research in the field can be technically impeccable, but without some common dreams, subfields will become so specialized that succinct and broad visionary synthesis is impossible. It is one thing to refine our technologies, and other to refine the questions we are able to ask. For this review, I will ask the following questions of each approach covered:

1) What is a good example of the approach? How does it work in practice?

2) What seeds lie at its historical foundation? What is its internal conceptual basis?

3) Where does it fall on the top-down/bottom-up spectrum?

4) How has it done in terms of its own goals and promises?

5) What has this contributed to strong AI?

6) Is it still going strong, as an approach?

Following the review, I will bring the threads back together, to consider how these approaches may

relate to each other within an overarching paradigm of strong AI, and compare this with how they may relate to each other without such a conceptual vision to unify their goals. Scientifically, these approaches are rarely at odds, and frequently complement each other, actually. If this review is useful as a synthesizing framework, then it will be a demonstration that the strong AI concept is still useful as a scientific impetus, and that its resurgence may be important for the future health of the field.

Classic AI: The Problem-Solving Paradigm

The General Problem Solver

In 1957, Newell, Simon, and Shaw created the General Problem Solver, or GPS. It was an ambitious project: theoretically, the system could solve any logic problem put to it. For its creators, it represented a major step towards the higher thought processes of strong AI – after all, runs the thinking in this paradigm, once a system performs logical operations, how far away can cognition be?

For the GPS, everything is a logic problem. Its world – that is, its problem space – consists of two things: objects, or symbols, and operators, or rules for manipulating those symbols. This metaphor – of objects and operators – is based on the idea that thought basically consists of symbols, and that those symbols are basically physical objects. That is, each symbol is a specific and unique physical state of the brain, a particular pattern of electrical activity (Newell et al, 1960).

The General Problem Solver works by building symbol structures, using the rules set by its Operators, that move towards and eventually solve its problem, or accomplish its goal. For example, proving mathematical theorems is based on the recombination of known mathematical axioms and theorems – mathematical objects - into unknown theorems – new structures – using the logical rules of established operators (ibid.)

GPS doesn't only solve simple problems with one or two steps, but complex ones constructed of a long series. For example, the GPS can play chess, in which it must construct a plan that includes many steps of operations. In chess, its goal is to place the opponents' king in a checkmate position. The objects of chess are the pieces; the operators are the rules. To achieve its goal, the GPS searches through many possible structures – that is, plans for the next series of moves – and chooses the structure that moves it closest to that goal. In order to find good moves, and to choose between them, the GPS must be programmed with chess-specific search criteria that include information about both positional advantages and the value of each piece given its position.

As the game develops, the GPS repeats this sequence of calculations repeatedly, navigating through its problem space to try to move closer to its goal. Internally, the system models this as a series of discrete "states" that, hopefully, lead to its "goal state" – that is, checkmate. For a program like the GPS, this series of states is called the "state space". The AI program considers many possible state spaces before choosing its next move, and tries to choose the one for which the potential outcomes seem to move it closest to its goal. The method it uses to find its best route is, therefore, called a "state space search strategy".

Searching Through the Problem Space

There are a number of ways to improve the effectiveness of a logic program like the GPS. In modeling real-world phenomenon, a better understanding of reality can yield better choices in what objects and operators are programmed into the model. Faster, more powerful computers mean that searches can be more exhaustive in less time, a phenomenon that will be covered in more detail in the Supercomputing section, below. But beyond these considerations, an effective state space search strategy is one of the key features of a working logic program. To understand the functioning of classic AI, at least a passing familiarity with search algorithms is necessary.

There are many metaphors that can be – and that have been – used to describe the nature of the problem space and the methodology a state space search strategy uses to navigate it, but one of the most ubiquitous is the tree metaphor. In this construction, the AI system begins at the tree's "trunk", and different branches represent different possible tacks or overall strategies. Within each of these general approaches, there are many specific moves or actions that could be taken. These are the leaves of the tree; leaves that are near the base of branches are obvious, likely possibilities, while far-fetched options are leaves at the branches' tips.

Programmers have various options for search algorithms – that is, formulae for how to proceed in a state space search strategy. A breadth-first search, for example, searches all of the most obvious possibilities, and then the next most obvious, and so on, while a depth-first search exhausts an entire "branch" before moving on to the next. Usually, more nuanced algorithms are more promising. Iterative Deepening Depth-First Search looks for a solution among all possibilities down to a specified depth before beginning the search again with an increased search depth, stopping when it finds a solution it deems "good enough". In Alpha-Beta Pruning, branches of the search tree are eliminated from future searches when they appear to be dead-ends.

All of these, and many more left unmentioned, are strategies for what amounts to a trial-and-error search. In other words, the system looks for the best solution it can find in a basically random way, as specified by its search algorithm. This is also called a *naïve* or *uninformed* search.

In another kind of state space search strategy, called an informed search strategy, the system exploits a priori knowledge about what kinds of solutions are expected in a given domain using sets of domain-specific search heuristics. Most chess computers, for example, are familiar with standard openings and common board arrangements. Upon encountering such arrangements, they can respond with a scripted, heuristic tactic.

The Problem-Solving Paradigm: History and Concepts

> *"The idea of an agent trying to change its internal state in order to come up with a sequence of actions as a possible solution to a problem or as a means to achieve a goal is deeply rooted in traditional AI and its view of the nature of intelligence."*
>
> - H.R. Ekbia (2008)

The Problem-Solving Paradigm, named for artificial agents like the GPS for which existence is an exercise in *navigating the problem space*, dominated thinking about Artificial Intelligence from the field's inception in the mid-1950's until its first winter in the early 1970's.

Over its long tenure, researchers from this school of thought tried to tackle almost every problem in AI. Obvious problem spaces, like chess, checkers, math proofs, and codes were early – and largely successful – efforts for the Paradigm, but later tasks included problems as diverse as machine translation, bipedal locomotion, perception, natural language processing, and nearly every other skill that has been ever been attempted in AI.

At the heart of the Problem-Solving Paradigm lie several assumptions about the nature of cognition. These sometimes explicit – and always implicit – concepts define the philosophy of classic AI:

The Syntactic Assumption: Symbols representing objects are entirely distinct from the experience of the objects in context. That is, a mental simulation of reality does not depend on any connection to or experience of actual reality in order to relate validly as a part of cognition.

The Algorithmic Assumption: Simple, linear, arithmetic math is all that is needed for mental operations.

The Disembodied Mind Assumption: The mind that deals with abstract thought can stand on its own, and doesn't need a body or sensorimotor skills to operate.

The Anthropocentric Assumption: Since symbolic logic is the seat of cognition, and since humans are the only creatures to use symbolic logic, artificial agents should be modeled after humans, or at least after human-grade expertise within a narrow domain.

The first three of these are the premises of classical logicism: the view that thinking consists of making logical inferences. The fourth is a natural extension of its precedents that is important for AI in particular. In essence, logicism represents the theory of reductionism applied to the mind itself, a movement that originally came out of Enlightenment-era rejection of consciousness as the province of the divine. Hobbes summarized logicism in 1651, with his famous assertion that reasoning – cognition – was "nothing more than reckoning". Around the same time, polymath philosopher Leibniz worked to develop a formal calculus of all human thought. He wrote, in a paper on the subject, "The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and where there are disputes among persons, we can simply say: Let us calculate without further ado, to see who is right." (Wiener, 1951). Kant, Hume, and other thinkers who followed built on these ideas, and in the following centuries, mathematicians like Gottlob Frege, Alfred North Whitehead, and Bertrand Russel more comprehensively formalized the concept (e.g. Frege, 1879; Whitehead & Russell, 1962). The premises of logicism were already solidly entrenched by

the time computer scientists like Alan Turing adapted them to computer science in the 1940s, and from there, it was a small step for the pioneers of AI – notably including the same Newell, Simon, and Shaw who created the GPS – to take them up as the theoretical foundations of the field (Ekbia, 2008, p. 23).

Even today, logicism remains an intellectual force. It seems to draw, particularly, on contemporary philosophers of the mind: for example, Jerry Fodor and Hilary Putnam (Horst, 2005). Another, philosopher Karl Popper, published a collection of his articles in with the unambiguous title, "All Life is Problem-Solving". Modern logicism is also called computationalism, and an especially popular manifestation of this is Newell and Simon's Physical Symbol System Hypothesis. In the words of the authors, "A physical symbol system has the necessary and sufficient means for general intelligent action." (Newell & Simon, 1976). Or, as Hubert Dreyfus later paraphrased, "The mind can be viewed as a device operating on bits of information according to formal rules." (Dreyfus, 1972).

To understand this premise, and precisely how it is tied to the assumptions of logicism, consider the following:

1) Formal rules are syntactic, meaning that the symbol is isolated from what it represents.

2) Formal rules are algorithmic, meaning that their operations are based in simple, linear, arithmetic math.

3) Formal rules are mechanical, meaning that they can be carried out by a simple system with no insight into the meaning of the operations beyond that needed to operate the algorithmic mathematics of the rules.

(Ekbia, 2008, p. 24-25)

For Newell and Simon, these are the "necessary and sufficient" conditions for cognition; for example, this is the way the human mind works. When Newell and Simon developed the General Problem Solver, a logic program based in precisely this thinking, it wasn't only a system with interesting deductive capabilities – it was a system that approximated, basically, human consciousness. Of course, Newell, Simon, and their cohort of problem-solving-based researchers knew that they were not there yet; but on the premises of logicism, they theorized that they had found and modeled the basic mechanism of cognition, and that fully strong AI could not be more than a few decades from that.

In that, they were wrong. Today, movement in the field is generally away from the assumptions and premises of logicism and the Physical Symbol System Hypothesis. In the sections to come, I'll cover that movement in detail, but first it bears reminding that we should be cautious in judgments we level at AI's pioneers. In retrospect, that they were mistaken is obvious, but how could they have known that at the time? In fact, what we know now is largely due to what we learned from their successes and their failures. Further, in the context of the spiritualism of consciousness and in the sheer

ambition that they dared, these founders of the field were still the intellectual rebels of their times. Even among the rational and scientific minds of their time they were rebels, since behaviorism still dominated psychology and theories based in symbols and internal logic are clearly not based in the observable behavior of good behaviorist science. To dare to look for intelligence in programs of constructed logical rules, to dare to found a visionary intellectual movement based not in a single established field, but a cross-disciplinary association of fundamentally radical thinkers, to dare to manufacture consciousness itself: all of these flew in the face of humility and religion alike. What details the pioneers of AI overlooked, how they fell short of their goals, must be considered in the light of exactly how grand the scope of those goals really was.

The Original Top-Down Models

Logic programs like the GPS aimed to model higher cognition exclusively with programming based in what the designers supposed were the formal rules of symbolic reasoning. To this way of thinking, if a program could emulate the rules by which humans reasoned and solved problems, then that program was unequivocally intelligent – with or without any deeper processes going on beneath. This is a philosophy that is well captured by the Turing Test: that is, in its most fundamental form, an assessment based on the premise that if a human judge cannot tell the difference between a machine's and a human's ability to exhibit intelligent behavior in a given domain, then the machine is intelligent in that domain.

During the years of the Problem-Solving Paradigm's ascendency, this framework was largely taken for granted. Emergence – that is, the concept that complex phenomena like cognition emerge out of simpler pieces like neurobiological connections – hadn't yet significantly entered into AI discourse, and the bottom-up school of thought didn't yet exist in strength. AI research during the Problem-Solving Paradigm years was basically synonymous with what we would not call top-down methodology: that is, if researchers wanted to model complex behavior, they searched for a set of logical rules that explained that behavior, and programmed those rules explicitly into a system like the GPS.

The Failure of Logic and the Lessons Therein

The General Problem Solver, and sister AIs from the Problem-Solving Paradigm, could play chess, prove mathematical theorems, solve puzzles, break codes, and much more. With an appropriate set of objects and operators, logic programs could theoretically attempt any problem posed in symbolic terms. All of this came on the footsteps of the computational revolution; when most people still hadn't seen a computer themselves, the first successes in AI were already springing up. At that pace, it's no wonder that the field's pioneers dreamt ambitious dreams.

Why, then, did the Problem-Solving Paradigm collapse? In 2007, Nils Nilsson laid out what he saw as the main grounds upon which the Problem-Solving Paradigm has been criticized. For him, the

critics of the Problem-Solving Paradigm fall into camps like advocates of non-symbolic processing, advocates of emergent or swarm-based intelligence, general rejecters of the computational model, and others who have largely theoretical disputes with logicism and its approach to cognition. While many of these are interesting theoretical questions, it was the trailing-off of early AI researchers' successes that led to the public's retreat – not theoretical squabbles. When Lighthill and other critics attacked AI research policy, they did so from a foundation of promises that researchers had definitively failed to live up to. With early and compelling successes, AI researchers had expected the rest of the strong AI puzzle to fall smoothly and rapidly into place. Why didn't it?

Since the early years of AI, a great deal of thought has gone into that very question, and many possible answers have emerged. In fact, most of the major approaches of AI can be conceptualized as emerging out of the various answers to which different researchers came. In the following, I'll review the main theories that followed the decline of Logicism, and what those concepts subsequently became.

Embodiment

One thing that AI's pioneers had no way to know was this: arithmetical operations on symbols are *actually very simple*. In terms of the amount of information being processed, even quite complex mathematical tasks are quite simple compared to the great variety of real-world situations that humans encounter every day. The ability to pour a glass of water, for example, requires a huge amount of declarative knowledge as one manages balance and fine motor control, processes very complex incoming perceptual data, and generally navigates a nuanced and ambiguous real world.

Humans may find mathematics and logic so difficult only because we have come very recently, evolutionarily speaking, to symbolic operations; logical thought is probably not the definitive pinnacle of cognition that was historically taken to be (Moravec, 1988). Early programs excelled at mastering logic, strategy, and mathematical theorems very rapidly because these operations are, objectively speaking, not very sophisticated. Researchers greatly overestimated their accomplishment because *humans* find logic, strategy, and mathematical theorems so difficult. This phenomenon later became known as Moravec's Paradox (ibid.), and out of it came the philosophy that it will be impossible for a computational system to perform many complex abstract tasks – including the development of a consciousness and the achievement of strong AI – without first mastering the more fundamental capabilities that humans have painstakingly evolved, especially sensorimotor processing and the possession of a physical body.

This emphasis on the necessity of sensorimotor mastery challenges logicism's Disembodied Mind assumption: that is, that the mind that deals with abstract thought can stand on its own, and doesn't need a body or sensorimotor skills to operate. Rule-based systems did very well for simple, artificial environments, like a simulated chess board, that were essentially graphical representations of mathematical operations. However, especially for early researchers, it turned out to be impossible for early logic programs to amass enough knowledge to perform even most simple tasks involving the

kinds of ambiguous data that humans deal with every day when moving, perceiving, and communicating.

Following this critique, some researchers began working on logic programs with the express intention of amassing exactly the huge database of declarative knowledge that might overcome this hurdle. Cyc, for example, was one of the most advanced question-and-answer AI systems in existence, for many years. Its creators have been growing its information database since the mid 1980's, allowing Cyc to informatively answer questions typed to it on a great variety of topics. As one would expect of a program designed around a philosophy of amassing tremendous knowledge, what distinguishes Cyc is not unique programming or an ability to learn, but its immense private database. However, as knowledge has become more and more interconnected with advancing internet technologies, other kinds of encyclopedic programs – like Watson, discussed in the section on DDAI, below – have overtaken Cyc. These programs are able to rely not on a private database but on the ability to interpret information that already exists – Watson, for example, is able to gather information from encyclopedias and internet resources in their existing form, and as a result, has the potential to outmatch Cyc by far.

For other researchers, the field of cybernetics seemed to be a more promising response to Moravec's Paradox than the work to overcome it by sheer force of declarative knowledge. The cybernetic approach emphasizes sensorimotor skills over logic, to try to build our way up to cognition by starting where nature did – with basic movement and perception. Not only is this a challenge to the Disembodied Mind, then, but it also repeals the Anthropocentric Assumption by rendering non-human creatures, even organisms as simple as ants and bacteria, as potentially informative models for AI.

This movement, which led to the rise of robotics and, more than any other school, came to replace the Problem-Solving Paradigm as AI's top dog, has also been called the Embodied movement for its emphasis on the physicality of artificially intelligent agents. Through robotics, the Embodied approach strongly influenced the entire diverse field of Nouvelle AI, detailed in the Nouvelle AI section below.

Symbolic Grounding

Another criticism of early work is that most rules-based AI lacked *symbolic grounding*: that is, the symbols the programs operated on were only abstract bits of mathematical information, and were not tied to perceptions and interactions with real objects. To advocates of this critique, the pure manipulation of symbols without some experience of what those symbols represent is so far removed from what actual organisms do as to be another thing entirely, and so not a fit model for intelligence (Nilsson, 2007). Another related critique is the argument that manipulation of symbols out-of-context is just mindless computation, and has nothing to do with thought or experience (ibid.)

This critique, in other words, is a direct challenge to the Syntactic Assumption of logicism: the assumption that symbols representing objects are entirely distinct from the experience of those objects

in context. For example, is the word “apple” equivalent to an apple itself, without even the memory of what an apple looks like or tastes like? Can a system that knows the word, but not the taste or image, not the context, really be intelligent about apples?

The robotics approach managed to generally address this criticism by building systems that actually interacted with real objects and people, but other researchers took another tack. Chef, a recipe-creating AI, was designed along the principles of Case-Based Reasoning. CBR, a research philosophy in the cognitive sciences, emphasizes the use of a set of experience records – analogous to memories – in order to learn from their generalities. Arguably, CBR challenges the Syntactic Assumption by tying rules to the actual experiences from which they're derived. Chef, specifically, considers a set of recipes and the variable successes and failures of that set. Also using its own knowledge about the principles of cooking, it creates original step-by-step food preparation instructions that are meant to avoid the less savory culinary outcomes, so to speak. In addition, Chef updates its understanding about cooking in general, as well as altering the specific recipes it's working on, based on its “cases" (Hammond, 1989).

Chef was fairly sophisticated for its time, but with the modern interconnectedness of information and advances in Bayesian mathematics – to be covered in more detail later – many contemporary DDAI programs are able to better accomplish similar kinds of tasks with much larger databases of information. Today, it's questionable whether Chef truly escaped the Syntactic Assumption at all; its operations may have been based in real information, but Chef itself certainly never touched a spatchula or cracked an egg.

Nonlinear Math

Problem-Solving Paradigm AI was based on the mathematical transformation and recombination of symbols. In other words, its programs relied upon simple, linear algebra; a reliance implying that logic programs could only model certain kinds of relationships and algorithms. Beginning especially with the rise of recursive connectionist networks in the 1980s, a number of mathematically intensive approaches have sought to model these kinds of nonlinear behaviors.

These approaches challenge the Algorithmic Assumption: that is, that simple, linear, arithmetic math is all that is needed for mental operations. Intensive math approaches developed along the principle that the operations and relationships within mental systems are in some ways beyond our current knowledge, and that therefore new kinds of mathematical representations may be needed to model these systems. Feedback relationships in dynamical and connectionist systems, statistically "fuzzy" relationships, and complex structural relationships like cause-and-effect or hierarchical organizations are only a few examples of real-world phenomena that fall beyond the ken of basic linear modeling techniques. From those limitations, systems and parallel processing mathematics, pattern-analyzing Data Driven Artificial Intelligence, and Bayesian graphical modeling arose as major AI approaches. Each of these will be covered in some detail in sections to follow.

Successes of the Problem-Solving Paradigm

Early research in AI wasn't only instructive for its limitations. Actually, its most striking successes were not in AI specifically, but in the field of cognitive science in general. AI projects, by their very nature, ground the theories upon which they are based. They are strong tests of ideas that may otherwise remain purely theoretical. If a theory of learning is entirely invalid, for example, then no amount of clever programming will render successful an AI based upon that theory. From early in AI history, computational models have informed and been informed by psychological theory.

For example, as the first problem-solving programs emerged in AI, psychological researchers were discovering evidence that humans used a very similar kind of symbol processing in their own advanced logical thought. Psychological models of reasoning began to emerge, detailing exactly what steps people used in order to work through logical problems. AI researchers were able to further test these models by building programs that used precisely the same strategies, and based upon the successes and failures of these programs, psychological researchers were then able to refine their theories and design follow-up research (Dreyfus, 1979). With this kind of computational modeling, called cognitive simulation, experts were able to articulate exactly what went on inside the logical process, exactly how the systems manipulated symbols. This collaboration, and the attention it brought to internal processes of the mind, brought about the decline of behaviorism – with its focus on observable behavior only – and the rise of a computational metaphor for cognition – that is, the rise of the field of cognitive science (ibid.)

Logicism Today

There's no doubt that the human mind uses symbolic logic, and that a strong AI based on the human mind will almost certainly be able to use it too. However, the idea that logic is the central, fundamental, or *only* pillar of thought has lost much popularity in the field.

That said, logical, rule-based AI is still a very visible dream of AI in the public mind.

Though declining among scientists, the Problem-Solving Paradigm still takes its large share of the cultural iconography of AI. Certainly, this is due in part to the disconnect between contemporary research in the field and popular culture representations of AI, leaving the public with an antiquated conceptual basis.

The programming of game AIs, from chess computers to simulated opponents in first-person-shooters, is also still predominantly based on rule-based problem-solving approaches. Though there have been forays into learning programs and emergent intelligence in game AI, this general orientation probably won't change anytime soon; as a commercial industry, game design companies have been primarily oriented towards "good enough" AI designs, and with enough computational power, rules-based approaches seem to work well most of the time.

In spite of the general trends, expert system and logic programming projects also persist within the research community. Remaining advocates of the Problem-Solving Paradigm have gathered under the banner of GOFAI: Good Old Fashioned Artificial Intelligence – a term coined by John Haugeland in 1985. For reasons I will explore in the following section, many of those who remain committed to logicism now look to the progress of computational hardware and supercomputing as the best hope for the success of their approach.

Supercomputers

The World's Biggest Brain

In 1943, World War 2 raged across the globe. In the USA, at the University of Pennsylvania's Moore School of Electrical Engineering, a secret military project began which would change the face of computing across the globe. Finally unveiled to the public in 1946, the Electronic Numerical Integrator and Computer – or ENIAC – was a 30-ton "giant brain" that could solve a mathematical calculation "... 1,000 times faster than it has ever been done before." (Kennedy, 1946). In other words, conceived and designed by University of Pennsylvania's John Mauchly and J. Presper Eckert and funded by the United States Army, the world's first general-purpose electronic computer was born.

ENIAC was first commissioned to calculate artillery firing tables for the Army's Ballistic Research Laboratory. To the modern eye familiar with complex computational calculations, this may sound like a mundane task, but prior to the ENIAC it was anything but simple. At that time, a trained technician with a desk calculator could compute a 60-second artillery trajectory in about 20 hours. State of the art technology of the time, an analog differential analyzer that was dedicated to these kinds of calculations, could produce the same result in 15 minutes. ENIAC, on the other hand, calculated a 60-second trajectory in 30-seconds: half the flight duration of the artillery itself.

In addition, unlike other computing machines, the ENIAC was not domain-limited: it enjoyed unprecedented mathematical flexibility, which meant that it could be – and soon was – applied to fields as diverse as calculations for wind tunnel designs, atomic energy, weather prediction, and much more. ENIAC was the most powerful and useful tool for major mathematical calculations in the world until other computers began to emerge in the early 1950s, and even those were just refinements of the ENIAC's design.

By today's standards, the machine was a gigantic beast and a slow one to boot, but it remains notable as the grandfather of modern computers. Its ability to make complex calculations in any domain was a feat largely made possible by its fully electronic design – unlike many of its predecessors, it relied almost entirely on electronic radio tubes, not on the mechanical or electromechanical moving parts which often resulted in more specialized supercomputation. Its general-purpose, cross-domain computability made the machine Turing-Complete – the first Turing-Complete computer in history.

ENIAC was a 30-ton, 1800 sq. ft. monster of a machine, containing 17,468 vacuum tubes, 7200 crystal diodes, 1,500 relays, 70,000 resistors, and 10,000 capacitors (Weik, 1955). Today, a system with the same capabilities can be simulated on a microchip the size of your fingernail: in fact, it already was, *fifteen years ago*. The University of Pennsylvania electrical engineering department – the same department that fifty years ago designed the original ENIAC – created a tiny duplicate in an evocative demonstration of the progress of computing. As the department reports, "...the ENIAC's 18,000

vacuum tubes and 170,000 resistors were modeled with 250,000 tiny transistors, mechanical switches were replaced with electronic ones, and digit and programming trunks were implemented as tiny metal lines interconnected through cross-point switches. The chip performs the same functions that its 30-ton predecessor pioneered 50 years ago." The designers even created a do-it-yourself kit, so that other organizations could put together a mini-ENIAC of their own (Spiegel, 1996).

The rule of supercomputing, and the premise of its contribution to strong AI, is simple, compelling, and obvious: that computational hardware gets better every day, and that with it, computers' capabilities are constantly becoming an order of magnitude faster and more powerful than they were before. Supercomputing, as I use the term in this section, refers to this phenomenon and its implications.

To one way of thinking, hardware is everything. The basis of supercomputing is not really in AI, but in information technology generally. For this reason, because this camp rarely makes claims that infringe on the theory space of other approaches, it is mostly seen as complementary with other movements, rather than in opposition to them. Nearly any AI project can benefit from faster, cheaper computation. As a result, supercomputing doesn't generally enter into the top-down/bottom-up debate.

Computation's Unambiguous Success

The hardware of computing has inarguably advanced at a stupendous rate. The speed and power of modern computers has, in a very real way, made modern society possible. A great number of the artifacts we depend on, from weather and natural disaster prediction to agricultural systems, from transportation systems to our entire banking economy, depend heavily on supercomputers. Their advancing power has enabled us reach for the stars, literally and figuratively. Computer hardware research has delivered on its promises, and more.

Most of the computer scientists and engineers who have provided advances in computer hardware never made any promises about AI in the first place. On its own terms, then, the computing industry is an unambiguous success, and continues to profoundly transform the world.

The question of its applications to strong AI is a separate one. In spite of efforts, strong AI remains one of the benchmarks that yet eludes the supercomputing industry.

The Church-Turing Thesis and Moore's Law

In the early 20th century several scholars including Alonzo Church, Alan Turing, and J.B. Rosser attempted to formalize an account of computability. These ideas came out of ponderings about determinism, a discourse about as old as theology itself. In other words, they wondered: given complete knowledge of a system's starting conditions and its set of operational laws, can we then make accurate predictions about its behavior and outcomes? Any system for which we can answer "yes" is a

deterministic system. For any deterministic system, we can – theoretically – mathematically model behavior: therefore, such a system is computable. It is not necessary that we actually have the knowledge or tools to model that behavior, only that it could theoretically be modeled if we did hold those tools. In Turing's words, "It was stated... 'a function is effectively calculable if its values can be found by some purely mechanical process. We may take this literally, understanding by a purely mechanical process one which could be carried out by a machine." (Turing, 1939).

For thousands of years, it has been debated whether or not complex systems like life or the human mind are deterministic and computable, a question that is intimately related to the pursuit of strong AI. After all, Computers are at their heart "computing machines"; in other words, they deal exclusively with calculations of computable models.

Any machine that can carry out mathematical computations generally is called "Turing-Complete", named for Alan Turing's foundational theories of computer science. Many early systems, and even some contemporary special-purpose supercomputers, only perform specialized functions in certain domains, and so are not Turing-Complete. For example, the eleven Colossus supercomputers used by British codebreakers during WWII were dedicated cryptanalytic systems; they were unable to complete most unrelated computational tasks, and so were not Turing-Complete like ENIAC, their contemporary. As recently as 2002, Gravity Pipe 6 may have been the fastest machine in the world for calculating astrophysics and molecular dynamics, but it could only operate in domains analogous to astrophysics and molecular dynamics, and so was not Turing-Complete. After the first decade or so of modern computing, Turing-Complete machines became the rule and dedicated hardware the exception.

Any machine that is Turing-Complete can theoretically work out any solvable mathematical problem, given enough time. In fact, that is all that a standard computer can do; all other computer functions are based on this one. Therefore, foundationally, all standard computers do the same thing: any Turing-Complete machine can theoretically model any other Turing-Complete machine. Theoretically – and only theoretically – the most primitive machine designed decades ago can mathematically simulate the functions of the most advanced computers today, given enough time. This is Turing-Equivalence; the observation that all Turing-Complete machines share Turing-Equivalence is the Church-Turing thesis.

What are the implications of the Church-Turing thesis for AI? Consider: If the mind is a computable system, then no matter how sophisticated that system is, any Turing Machine should theoretically be able to simulate it. It is simply a matter of two prerequisites, which I will refer to herein as the *barrier of mental knowledge* and the *barrier of computational hardware*. The first involves the identification and programming of a model with conditions and processes that enable intelligence. The second involves developing a computer that can run such a model at a reasonable – that is, observable – rate. A primarily supercomputer-oriented approach to AI is one which more or less takes for granted some method of overcoming the barrier of mental knowledge – the work that other approaches to AI emphasize – and instead focuses on development towards overcoming the barrier of

computational hardware.

Luckily, better hardware seems to develop at an unceasing pace. As Gordon Moore made famous with his 1965 observation, just four years after the first planar integrated circuit was created, improvements in computer hardware weren't only steady, they were *exponential*. In his original paper, Moore observed an exponential growth in the number of transistors per integrated circuit – specifically, they doubled in density about every two years – and predicted that this trend would continue (Moore, 1965). In spite of repeated predictions that this could not continue forever, Moore has so far proven correct year after year, and the phenomenon has been coined Moore's Law. Advances in hardware are still steady, perhaps because we apply each new year's advances in hardware to innovations in the years to come, and Moore's Law could continue well into future decades.

In other words, today's supercomputers are tomorrow's laptops, which in turn will enable tomorrow's supercomputers to be exponentially more powerful than today's. For the supercomputing approach to AI, the theory is that eventually, this advancement will result in computers so powerful that they are able to complete any mental task at least as well as a human can through sheer computational speed and power. Is this projection realistic? And if not – if computational power will not unilaterally overcome the challenges of strong AI – then what exactly does progress in supercomputing have to do with the strong AI endeavor?

Computability and Parallelism

It is unknown whether the Church-Turing thesis was fully correct, and it is possible that biological intelligence is based on mathematics that not just any Turing Machine can simulate.

For example, computers typically run on a single central processing unit (CPU). In essence, a single CPU can only do one thing at a time, with some variation as to how complex this thing can be. Specifically, scalar processors can operate a single instruction for a single item in a data set, while vector processors can operate a single instruction on multiple items at once. Modern computers create the illusion of being able to multitask via concurrency: that is, by shifting between many tasks very quickly (Barney, 2011).

The human brain, however, doesn't rely on concurrency to multitask: it engages in genuine parallel processing. In other words, by having many components that can each act autonomously, it is able to actually do many things at once, not just to fake it. Some computers are able to do this, as well, by hooking up networks of many CPUs. In fact, this kind of networking is the basis of modern supercomputing. For example, the IBM Roadrunner – currently in the world's top ten supercomputers in terms of speed and power – contains 12,960 IBM PowerXCell CPUs and 6,480 AMD Opteron dual-core processors (Koch, 2008).

It is feasible that this kind of massive parallel processing manifests certain systemic properties that are impossible for a single CPU to manifest. The hardware of the brain is uniquely different than the

hardware of a basic Turing Machine, and it may use mechanisms like parallel processing that such a machine cannot accurately emulate. This would not indicate that the brain wasn't computable, only that it would require hardware with abilities beyond what Turing envisioned. In this case, some of the assumptions at the heart of Turing-Completeness will need to be reconsidered. It is still commonly believed that Turing was correct and that even characteristics like parallelism can be accurately simulated by any Turing-Complete machine. However, even if cognition does turn out to be a computable system, it wouldn't necessarily follow that increased computational power is all that is required to overcome other barriers to strong AI. To understand this better, consider supercomputation's success in the domain of chess; that success taught us a great deal about what a raw-power approach to AI can do, and a great deal more about what it can't.

Supercomputing and the Chessboard

The chessboard was one of the earliest arenas for efforts in AI, but though the earliest chess-playing programs were developed as early as the 1950s, they made very slow moves and were easy to defeat (Wall, 2011). Chess computers became more and more sophisticated in the following decades, but it wasn't until 1996 that IBM supercomputer Deep Blue outscored world chess champion Kasparov in a series of matches: the first computer to outperform the best humanity had to offer on the game board. Interestingly, Deep Blue still used basically the same chess-playing strategy as chess programs from decades before. Though computer knowledge of chess had advanced and search algorithms had improved, Deep Blue's methodology of a state space search strategy was fundamentally no different than that used by the General Problem Solver decades before, and by the first dedicated chess computers before that. How, then, did Deep Blue overcome Kasparov when none of its predecessors could? Deep Blue, or more accurately the IBM project team that designed it, took its victory not by understanding chess better than Kasparov did, but by utilizing the overwhelming processing power of a supercomputer to search through a hundred million possible moves per second and finding the state space most favorable for its victory (Hsu, 2002).

Before Deep Blue, it was commonly believed that the ability to play chess well was probably indicative of the ability to think creatively: in other words, people assumed that Chess was essentially an AI-Complete problem, solvable only by a strong AI. After all, chess had been held up as a hallmark of intelligent thought in humans for many years (Hofstadter, 2001). For a supercomputing approach, Deep Blue was a powerful success story. That a computer could surpass the best of humanity in a domain as complex as chess suggested to many that it was only a matter of time before sheer computational power overcame obstacles in other domains like speech and insight. Someday soon, to this way of thinking, we would necessarily develop computers with even greater speed and power than the human brain, and strong AI would lay only a step away (Kurzweil, 2005).

Unfortunately, such conclusions were premature. The success of supercomputing, problem-solving AI at chess didn't prove that strong AI was at our doorstep; it proved that like logic and math, chess was not what we once thought it was. Chess, it turns out, can be broken down as the

representation of a formal system of symbols: a set of mathematical theorems (Haugeland, 1985). For humans, chess skill is at least somewhat related to the ability to think generally – at least in terms of logic and problem-solving. For computers, it's not (Hofstadter, 1985).

At this point it seems very unlikely that processing power alone will overcome deficiencies in the design of strong AI. After all, even Deep Blue was not operating only on power: it was operating from the Problem-Solving Paradigm, and with an IBM team that included a chess grandmaster to assist in its strategies.

Supercomputing improves the speed at which machines process information. Part of its compelling story, then, may be based in the fact that when we observe humans, we often equate speed of processing with intelligence. When we say a person is "quick-witted", we imply intelligence; when we say a person is "slow-thinking", we imply a lack thereof. However, this lay intuition of what intelligence is simplistic at best. In fact, processing speed has little to do with the quality or complexity of thought that a system can manifest; consider Turing Equivalence: even a very slow Turing Machine can, given infinite time, perform just as complex calculations as the most complex. In corollary, faster processing speed doesn't automatically grant new abilities to a system.

In other words, a pure supercomputing argument ignores the *barrier of mental knowledge* mentioned previously, or at least assumes that a computer can overcome it with enough brute force at its disposal. In the same way that Deep Blue's ability to perform massive searches – a "counting out" strategy, as coined by Dreyfus (1992) – eventually overcame Kasparov's ability to reason creatively – a "zeroing in" strategy (ibid), the fundamental tenet of a strong supercomputing perspective on AI is that the application of raw computational power must eventually overmatch humans in other tasks like natural language performance and sensory processing as well.

Instead of viewing the brain as a simple computational machine, it is increasingly validated as a constructive learning system, in which the operations of the system change qualitatively as the system learns (Quartz & Sejnowski 1997). The computer metaphor for this would be to liken the brain to a machine that programs and improves its own software. It is likely that the key to strong AI may lie in finding a way to model this, rather than in modeling computational power alone. We remain only near the beginning of our quest towards identifying the starting conditions and rules of mental life.

That said, a home office laptop of today is far more powerful than a supercomputer of twenty years ago, and this progress does have a profound effect on AI research. By increasing the accessibility and affordability of powerful machines, the access of researchers to cutting edge work has become more and more widespread, which can only increase the pace of developments in AI.

Genetic Algorithms

Genetic algorithms are an approach to AI that relies of brute-force computation to generate innumerable examples of random strings of code, from which an AI program can select the ones that

seem to be most useful and incorporate those into its program. They also provide one of the few possible exceptions to the rule that, on the whole, the role of supercomputing remains an ancillary and enabling one for strong AI research. For at least some proponents of this approach, genetic algorithms offer a strategy for using ready advances in computational hardware to circumvent the *barrier of mental knowledge* that remains problematic. Genetic algorithms can be considered an attempt at manufacturing a digital analog to cognitive evolution, wherein random bits of code replace random mutations to biological genetic information. Theoretically, with enough raw computing power, a system can generate billions or trillions of alternative cognitive algorithms, and throw them into competition via algorithmic selection mechanisms, so that the ones which most effectively achieve their target tasks emerge as the choice cognitive strategies of the artificial agent (Holland, 1975).

To those who hope for strong AI in the near future, genetic algorithms offer a very attractive approach. They seem to depend predominantly on computational power alone, and computational power is something that we already understand well and will almost certainly have more and more of in years to come. This has led to some proponents making what are probably unrealistic claims about where genetic algorithms will lead us and how quickly their progress will be.

Ray Kurzweil, a popular figure in public discourse on AI, has come out strongly in favor of the union of supercomputing and genetic algorithms. He has estimated that by 2029, give or take a year or two, we will almost certainly have achieved strong AI: a claim for which Kurzweil has received heavy criticism by the scientific community. In his words: "All that is needed to solve a surprisingly wide range of intelligent problems is exactly this: simple method combined with heavy doses of computation..." (Kurzweil, 1999). The "simple methods" he refers to are primarily genetic algorithms.

Unfortunately, though a potentially powerful tool, genetic algorithm programs still rely on programmers who understand the function of intelligent behavior and cognition well enough to program in the criteria by which the program selects "successful" algorithms (Mitchell, 1996). In this way, the genetic algorithm approach does not escape the first problem of AI: the *barrier of mental knowledge*. Most scientific proponents of Genetic Algorithms understand this, and are, in fact, highly critical of those like Kurzweil who seem too blindly optimistic about the approach's near-future potential. In the words of Hofstadter, himself one of the originators of the genetic algorithm approach, the claims that "intelligence would simply grow automatically out of genetic algorithms, much as a tomato plant automatically grows from a seed in a garden, [are] ungrounded fantasies." (Ekbia, 2008).

Supercomputing Today

Every day, advances in supercomputing continue their incredible progress, for reasons that usually have nothing to do with strong AI. The engineers developing supercomputation these days are rarely concerned with its applications for strong AI, and most serious proponents of strong AI generally agree that supercomputation alone will yield few breakthroughs for the field. However, all of that said,

progress in strong AI is still intimately linked to progress in computational hardware; the enabling potential of Moore's exponential growth cannot be ignored. The hardware of computation is the skeleton upon which all AI models must be built, and to conceive of AI's future, it is necessary to understand something of the shape and the growth of that skeleton

In 2011, there are thousands of supercomputers around the world. The fastest is the Tianhe-1A, a Chinese system with a peak computing rate of over 2.5 petaFLOPS – that is, (2.5)1015 floating point operations per second, an abstract measure of a system's performance speed –, nearly 70% faster than its closest competitor: the Cray Jaguar which operates at 1.75 petaFLOPS (Stone & Xin, 2010).

However, these lions of the supercomputing world are not actually the most powerful artificial systems in the world. That honor goes to large-scale distributed computing systems, specifically ones that borrow computing power from hundreds of thousands of personal computers that are all connected to servers that, in turn, are connected to each other.

A distributed network borrows power from the connected computers of volunteer users when that power isn't being used. For example, the BOINC network draws power from over 480,000 active computers at any given time, giving it over 5.5 petaFLOPS to apply towards various distributed computing projects (Folding@home: OS Statistics, Stanford University, retrieved 2011-05-28). Another example, MilkyWay@home runs on the BOINC system – using about 33,000 active computers – in order to generate highly accurate, up-to-date 3D models of stellar streams near to the Milky Way galaxy (BOINCstats: BOINC Combined, BOINC, retrieved 2011-05-28).

In fact, the internet itself can arguably be viewed as a giant distributed computing system; Google's computational cluster, for example, has recently been estimated as able to run between 20 to 100 petaFLOPS on over 450,000 servers (Markoff & Hensell, 2006). In a future where the interconnectedness of computational systems becomes more and more ubiquitous, the already permeable divisions between computers may become even less important. The supercomputational power of the future probably lies increasingly in distributed networks that span the globe.

With Moore's law, we can also expect to see a continuation of the miniaturization of microchip technology. It probably won't be long before microchip manufacturing becomes a nanotechnological task wherein microchip factory equipment operates at a level as small and smaller than the human eye can see. Nor will progress necessarily slow when we do eventually exhaust our miniaturization options: one direction for next-generation development lies in quantum computing. Quantum computing is a nascent, largely theoretical domain involving the use of quantum rather than Newtonian mathematics as the root of computing. Quantum computing, which has been demonstrated as theoretically possible but which still faces many barriers, has the potential to exponentially outstrip traditional computing in power and speed (DiVincenzo, 1995). Results in quantum computing are still decades from fruition, but they may eventually lay the foundations for the future of computers.

At the moment, perhaps the most extensive concerted effort in supercomputing research belongs

to the Blue Gene Project, a collaboration between IBM and a number of other organizations with various visions and strategies. With this effort, IBM aims to create the next several generations of supercomputers. For example, the Sequoia, scheduled to go online in 2011, will operate at 20 petaFLOPS: the fastest in the world once it's complete. It will primarily be used for nuclear simulations, as well as for other scientific purposes (Seager, 2009).

Another project, the Cyclops64, focuses on miniaturization rather than power. Its designers term it the first "supercomputer on a chip". The Cyclops64 will use cellular architecture to perform at 80 gigaFLOPS per chip. To put this into perspective, the NVIDIA GeForce 8800 X, with processing chips fairly typical of a contemporary home computer graphics card (GPU), operates at about 518.4 gigaFLOPS on 128 processing chips, or at about 4 gigaFLOPS per chip. In other words, the Cyclops 64 chips will be about fifteen to twenty times faster than contemporary home computer graphics chips, or fifteen to twenty times smaller for the same functionality. If instead compared to a typical CPU – which processes fewer FLOPS than a GPU of the same grade – the discrepancy becomes even more striking. In terms of chip miniaturization, the Cyclops64 will jump a generation or two ahead of current advances (del Cuvillo et al, 2004).

At this rate, it won't be long until every home computer performs at the petaFLOPS level, and soon after that even petaFLOPS will be a measurement of the past. If Moore's Law holds, then by 2030, we'll have well surpassed petaFLOPS and even be saying goodbye to exaFLOPS, the next order of measurement. Extrapolating current trends, we will see our first zettaFLOP supercomputers emerge in less than twenty years ("IDF: Intel says Moore's Law holds until 2029". Heise Online. Retrieved 2008-04-04). Outside of its applications in strong AI research, that level of computational power may enable us to achieve such feats as the fully accurate modeling of weather systems up to two weeks in advance (DeBenedictis, 2005). Within strong AI, such computational power has the potential to unlock many avenues of research, like massive neural nets analogous to the structure of the human brain. I will discuss this, and other examples of the kinds of new technologies that supercomputation may enable, in the chapters to come.

In general, the great successes of supercomputing in AI have never really been successes from supercomputing alone. Instead, they have been collaborations between innovative theories of cognition or task performance and the advances in computing that have enabled those theories to be implemented. Even Deep Blue involved human insights into the process of chess performance; even genetic algorithms require human understanding to create selection criteria.

Unfortunately, increases in the raw processing power of computational systems are the most easily understood aspect of their development. The public can easily grasp simple measurements of processing power, and scientists and journalists can easily report them in sound bites and short articles. As a result, progress in computational hardware gains much of the attention for successes that occur from qualitatively new approaches when in fact, as a rule, supercomputation has only performed an enabling – though a powerfully enabling – role.

Connectionism and Mathematical Modeling

Modeling the Human Brain

Connectionism and mathematical modeling, perhaps more than any other approach, have relied on these kinds of advances in computational power. The Blue Brain project is an ideal example. A collaboration with IBM's Blue Gene effort and with the Swiss government, the project rests – perhaps fittingly – on the same shore of Lake Geneva where Mary Shelley wrote Frankenstein. The project's director, Henry Markram, ultimately aims to model the entire human brain. Markram believes fundamentally in the obtainability of the project goals: in December 2006, he reported that he had successfully simulated a rat's entire neocortical column, and that the human cerebral cortex wouldn't be far behind (Witchalls, 2007).

Markram started with the relatively simple neocortical column of a rat, an intermediate structure within the anatomy of the brain chosen for its role as the smallest functional neural unit – much larger than a neuron and much smaller than a lobe or cortex. He plans to work his way both downwards and upwards from there: down into the details of molecular neurology and up into the workings of the entire brain. Currently, having simulated the lone column, Markram hopes to better understand the column's functions within the brain so that he can streamline its simulation and more easily allow the parallel simulation of millions of connected columns.

The Blue Brain team has managed to simulate the neocortical column of a rat, but a human's neocortical column contains about ten times as many neurons, and the human cerebral cortex is made up of something in the ballpark of two million such columns. To the project's way of thinking, each neuron will require the computational equivalent of about one contemporary laptop to simulate. This may seem insurmountable, but in Markram's view, tomorrow's supercomputation will be up to the challenge (Michio, 2011). "It is not impossible to build a human brain and we can do it in 10 years," Markram asserted, at a TED conference in Oxford. Nor does he think the project will be stop at descriptive simulation. As he said in a BBC World Service interview, "If we build it correctly it should speak and have an intelligence and behave very much as a human does." (Fildes, 2009).

These estimates should probably be interpreted with considerable skepticism. Many scholars have criticized Markram's claims as presumptuous, ungrounded promises. For example, as the Blue Brain project was first launched, Markram likened its ambitions to those of the Human Genome Project: a bit of self-promotion that many of his colleagues found grandiose and ridiculous (Lehrer, 2011). Computational neuroscientist Sejnowkski described Blue Brain as "bound to fail," since at this point, we don't understand cognition well enough to know specifically what its modeling requires (ibid.) However, disregarding the damage that may be done by overoptimistic claims, the scope of what the Blue Brain project is attempting is admirable and, with given ongoing access to computational

resources, it may teach us a great deal about both computer science and the nature of the brain.

Advances in distributed processing and neural nets have shown us that a great deal of computational power can be generated from simply having a massive, interconnected information-processing network. But is massive interconnectivity and parallel processing the key to strong AI? Can emergent, general intelligence – as a strong theory of connectionism would posit – be generated automatically from any interconnected system with enough nodes in its network? This is one central question raised by efforts like Markram's Blue Brain Project.

Dynamical Systems: Feedback and Recursion

Most mathematics used by researchers is linear in nature: that is, functions apply evenly to individual data points and don't change their application over time. Linear mathematics is the basis of most contemporary research because it is far easier to solve and model than nonlinear mathematics. In reality, however, very few complex systems operate on linear principles. For example, interactions that are any more complex than one-way cause : effect relationships require nonlinear modeling – also called *dynamical systems mathematics*.

A very simple example of dynamical systems mathematics applies in the case of population growth: as more animals are born into a population, that population gains reproductive potential, so even more individuals can be born more quickly. This is a simple nonlinear exponential growth curve, also known as a positive feedback mechanism, wherein a change function accelerates as its acts upon a system. In natural populations, this accelerating trend of "population explosion" continues for as long as pressures like food shortage and increasing predation are not prohibitive. However, as population increases, food shortages and predation become more and more prevalent, eventually reversing the accelerative function. This is a negative feedback mechanism, a stabilizing function that acts to mitigate systemic change and preserve the status quo.

Both positive feedback and negative feedback are mechanisms that are only found in dynamical systems: that is, systems that are *recurrent* in nature. In recurrent systems, causes and effects form repeating loops across time. Take, for example, a system in which factor A has an effect on factor B, and factor B has an effect on factor C. As it stands, this is a linear system – it is not recurrent. If, however, C also has an effect on A, then recurrence is now in effect, and so systemic behaviors like feedback can occur. The study of the mathematics of such systems, when the mathematics is not reduced to linear heuristics, has been called cybernetics, systems theory, or dynamical systems theory.

Consider another example of a dynamic system at work: a student learning vocabulary and concepts in a new language or domain of knowledge. At first, vocabulary growth might be slow, as the student has few existing words and concepts with which to help her grasp the new ideas. However, as time goes on, the student is able to recruit the vocabulary she's been learning to increase her pace of mastery. This function is analogous to the one at work in population growth: another positive feedback mechanism. This kind of function, and a plethora of other dynamic behaviors – some of

which are understood better than others – are constantly at work all throughout human behavior, cognition, and neurobiology. For example, the cerebral cortex is a massively recurrent network: one in which cognitive activation can spread and echo back upon itself.

These kinds of systemic behaviors seem to be the rule in the physical universe, not the exception. However, it is impossible to model them with classical, linear mathematics. As a result of mathematical limitations, the earliest connectionist models were non-recursive, linear models. In 1957, scholars at Cornell developed the first and simplest artificial neural networks, perceptrons, which simulated the process of neural activation but without recursion, and so without dynamical behavior (Rosenblatt, 1957). In other words, the perceptron's network only went A -> B -> C; it did not loop. Because the later effects of the network did not in turn act upon any of the earlier causes – in other words, because factors C and B could never affect factor A – the perceptron and systems like it were called "feedforward" networks. Perceptrons, though initially promising, were not able to live up to the hopes of their creators, and the field stagnated.

It wasn't until the 1980's, with advances in both computing and mathematics, that connectionism began to revitalize with the first recurrent artificial neural networks. These systems exhibited "directed cycles": fully recurrent loops within their nodes. Once developed, recurrent artificial networks showed many of the capabilities that feedforward networks did not, pumping life back into field (Rojas, 1996). Recurrence, it seemed, was the central feature of a successful artificial network – also called a parallel distributed processing (PDP) for the parallel – that is, simultaneously occurring – nature of spreading activation, and its distribution across many nodes within the network (McClelland & Rumelhart, 1987).

Connectionism, then, began as an effort to model the actual way that the brain uses a huge interconnected network of nodes to process information. However, in time, the recursive nature of these interconnected networks forged a close alliance between connectionist theory and cutting-edge, non-linear, systemic mathematics. Today, progress in connectionist architecture is intimately bound up with progress in the mathematics of dynamical systems (Rojas, 1996).

Connectionist techniques are not only used to model neurobiology. Other researchers have found them to be just as effective in simulating behavioral or cognitive phenomena, especially within the domain of language acquisition. Plunkett et al (1992), for example, developed a connectionist-based model for language acquisition that not only mapped words to their referents more successfully than its Problem-Solving Paradigm-based competitors, but that also exhibited language mapping errors similar to those of actual children, which may suggest a similarity between the mechanisms at work in this system and those that human language learners use.

Connectionist models are essentially emergentist, in much the same way that many projects within the Nouvelle AI approach tend to be. The emergentist philosophy is based on the synergy of the whole – on the idea that complex phenomena like consciousness rise out of the interacting relationships of a whole system and so cannot be understood as component pieces. By the nature of their emergentist

techniques, connectionist researchers lay out the starting conditions of an artificial Mathematical Models of Systems

Connectionism came from the point at which neurobiologists needed new mathematics to progress their science. Mathematical modeling techniques, on the other hand, stretch back much further. They predate psychology, predate even empirical science; probably, the first mathematical modeling of systems was performed by ancient astronomers as far back as the Babylonians, trying to predict and understand the movements of celestial bodies (Aaboe, 1974).

Early scientific psychologists worked to identify the mathematics at work in human neural network or other whole system and then watch to see what happens. This emphasis is distinct from setting out with a specific problem-solving goal within specific parameters, the kind of approach we are used to seeing from the problem-solving and supercomputing paradigms. Connectionists have used the term "tinkering" to refer to their more open-ended, exploratory designs.

For some researchers within the connectionist community, mostly neuroscientists, this bottom-up methodology has developed further into a scientific ethos. This "neuromorphic" perspective only sees value in connectionist models of the brain, like the Blue Brain project, and discounts connectionist models that work at the cognitive level like the language acquisition system mentioned above. This has been criticized as overly reductionistic (Geake, 2008), in part because its proponents often seem to emphasize an approach to strong AI that attends only to the science of properly constructing the hardware of the brain, not at all in the science of understanding the algorithms of the mind.

Behaviorist models of mind and cognition are originally based on animal responses (Skinner, 1938). Human behavior and cognition, however, was so complex that behaviorists could only make headway by radically simplifying its operations within their models. Since the Skinnerian era, psychological models have become steadily more complex in both theory and mathematics, such that today, the technique of mathematically modeling complex psychological and social phenomena forms an extensive sub-field of its own.

In most approaches to AI, researchers set out with explicit goals of, one way or another, creating artificial intelligent agents. Mathematical modeling, however, is not fundamentally an AI approach, but rather a scientific approach to understanding the world – a powerful tool for making particular theoretical assumptions explicit, and exploring their implications. Its ascendency in the mind sciences considerably predates most of AI's history.

Simultaneous to developments in the Problem-Solving Paradigm, psychological scientists were breaking out of the behaviorist mold to perform experiments on how humans solved problems and logically reasoned. For this, new math was needed, as well as new theories. The empirical, step-by-step analysis of these cognitive processes laid the foundation for the development of mathematical tools as a major piece of psychological research. This was not only the foundation of mathematical models of the mind – a discipline also called "cognitive simulation" – but also the foundation of cognitive

science more generally (Dreyfus, 1979). For none of these original influences on mathematical modeling, though, was the goal the *creation* of intelligent agents: the goal was only to better understand how intelligent agents worked. To this day, mathematical modeling researchers generally approach their science with the goal of accurately describing existing phenomena by using computational programs: a nuanced but real difference from the many AI researchers who model existing phenomena towards the explicit goal of understanding how to create more sophisticated artificial agents.

In addition, as the study of the nervous system became a real possibility, the mathematics for modeling neurological behavior – and eventually its relationship to thought and action – were in high demand. Early advances in the study of neurobiology codeveloped with the discovery of the mathematical algorithms by which neurobiology functioned (Griffith, 1971). Indeed, one way we can measure our contemporary understanding of the brain is to ask the question: For what behaviors have we – and for what have we not – identified functional algorithms?

Today, mathematics influences neuroscience and, in turn, neuroscience inspires new research in mathematics. For example, mathematical researchers have worked with neuroscientists to identify the mathematical bases for the neural network dysfunction responsible for Parkinson's disease, so as to design better treatments (Rosenblum & Pilovsky, 2004). It is because of this intimate bond between connectionist neuroscience and mathematical modeling that I've grouped the two approaches together here.

Bayes Nets/Graphical Models

One critique that can be raised about many approaches to AI, especially to rules-based expert programs – like Deep Blue for chess or Chef for cooking recipes – is that they overemphasize mastery in a specific domain and are not concerned enough with the task of *learning how to think*. The more advanced systems from the rules-based paradigms are programmed to learn, certainly, but only within very narrow domains. For example, a rules-based language-learning system is programmed to use cross-situational occurrences to infer new vocabulary (e.g. Siskind, 1996), but is extremely limited by its inability to organize that knowledge or to learn about – and learn from – grammatical organization.

This distinction between acquiring content and acquiring rules – between knowledge and grammar – is central to understanding new Bayesian approaches in AI learning systems. In part due to mathematical modeling limitations – though sometimes that is only the mathematical limitations of investigators – both psychologists and AI researchers have been slow to move towards theories of learning that embrace *grammar acquisition* as the rule for the development of complex cognition: that is, that focus on dynamically changing organizational strategies as the key to domain mastery (Gopnik et al, 2010). Even when artificial agents are programmed to learn, these learning strategies are typically simple and constant. In other words, agents rarely learn new ways of learning.

This is problematic, because what makes complex abstraction and reasoning possible is the way

that knowledge symbols are related to each other by a variety of organizational structures. As Poincare penned a century past (1905), "The aim of science is not things themselves, as the dogmatists in their simplicity imagine, but the relations between things; outside those relations there is no reality knowable." His observation applies not only to formal science, but to the ubiquitous and probably more important science of everyday learning and thought. To date, most computational models of learning have failed to take this into account. Researchers have sometimes equipped their programs with rules of learning, but until recently it was very rare that programs were equipped with rules by which to learn about learning.

In human children and infants, research increasingly points to the importance of the development of organizational relationship strategies as a key to developing cognition. Before they can even articulate a sentence, infants are already internalizing the structures that organize the relationships of its objects – linear orders, tree hierarchies, causal relationships, and many more – and mastering the organizational algorithms for these structures (Gopnik & Meltzoff, 1997). Without the ability to learn the rules and structures of learning, even the most sophisticated rule-based language learner will probably fall short of a human child in performance – and that's still ignoring the importance of attendance to social cues (Baldwin & Moses 2001 for a review). A child's ability to master the specific organizational structures and algorithms of its native language enables it to learn and apply both vocabulary and grammar with far greater fluency than any pre-programmed rule-based word learner. It is through flexibility and the ability to learn about relationships that humans deal with the ambiguities of their environments. There are rules of operation within language, certainly, but superseding them is a much more important rule: that language agents must constantly and adaptively master the ever-shifting rules of language operation – that they must engage in an ongoing search for better ways to organize new and old information – that they must constantly learn how to learn.

Some researchers took the first major step towards models of this kind when they applied Bayesian mathematics to the problem (e.g. Frank, Goodman, & Tenenbaum, 2009; Kemp & Tenenbaum, 2008). A Bayesian network adapts the function of prior probabilities into a statistical model, also called a directed acyclic graphical model. Such a model's ability to use prior knowledge in its operation means that, unlike most statistical models, it is not limited only to predicting correlational relationships. Instead, it can specify other organizational algorithms – most famously, causal structures – when fitting the data to its model. In layman's terms, a Bayesian network can organize statistical information into a variety of conceptual structures, something that other statistical models cannot do. Obviously, the discovery that a relatively simple statistical technique can enable a system to organize evidence into complex structures of concepts is very meaningful for AI research, and has been leapt upon by some researchers trying to model intelligence. Psycholinguists and modelers of language acquisition have been especially intrigued by the applications of Bayesian nets.

Most typically, it works like this: Bayesian researchers pre-program an organizational structure of word learning into the network based on their beliefs of how that learning occurs in the human system. In other words, researchers specify the learning algorithms, or strategies, that they think are at work for human language-learners within the computational system, and then they set that system

loose on a simulation of a real word-learning environment. This is distinct from most word-learning programs which rely only on co-occurrence of objects and referents rather than on grammatical strategies – the ability to use different kinds of grammatical structures to assist word learning represents movements towards AI which can incorporate true grammar learning. Essentially, these models are equipped with learning constraints that lead them to expect certain kinds of relationships rather than others. These kinds of Bayesian word learners can even be set up to use social cues, like speaker gaze, in their mapping decisions (Kemp & Tenenbaum, 2008).

This kind of model, one that is equipped with a specific organizational structure that fits the domain of the data, has shown advantages over other models on word learning and other tasks (e.g. Frank, Goodman, & Tenenbaum, 2009; Kemp & Tenenbaum, 2008). This is not surprising, given that learning constraints in terms of expected forms of relationships probably explain a substantial piece of how children are able to map new words, concepts, and organizations so rapidly and with relatively sparse evidence (Tenenbaum, Griffiths, & Kemp, 2006). However, these kinds of artificial agents don't learn new organizational structures – they don't acquire grammar – they are only pre-programmed with more sophisticated organizational structures than their predecessors.

In designing AI that is modeled after human cognition, the successes of such programs might seem encouraging to a nativist, rule-based view that suggests innate organizational structures for different kinds of information (e.g. Atran, 1998.) Chomsky (1980) words this argument directly: "the belief that various systems of mind are organized along quite different principles leads to the natural conclusion that these systems are intrinsically determined, not simply the result of common mechanisms of learning or growth." Pure learning efforts to dispute this view, like recent neural net learning approaches (Rogers & McClelland, 2004) have probably suffered some performance impairment in part for their monomodal organizational strategy – that is, organization of data by simple, linear correlation (Geman, Bienenstock, & Doursat, 1992).

Here, the strong AI concept may shed some light on the most promising future directions for designers of artificial learning agents. It seems increasingly plausible that the best model for such agents, rather than being nativist accounts or learning accounts with monomodal organizations, will be one that not only *uses* various organizational structures likes the Bayesian learning systems described above, but that continuously *creates* new organizational structures as part of functioning.

Certainly, among human learners, even fully developed adults continue to organize new data in a variety of ways, and make organizational choices based on an estimation of what strategy will best fit the information they are dealing with (Novick & Hurley, 2001). For children, it's conceivable – as nativist proponents would argue – that the cognitive organization of areas within certain domains is innate; but for most of what humans learn, research findings have pointed to a developmental process in which children gradually discover the nuances of structure (Rosch, 1978; Markman, 1989; Schultz & Vogel, 2004).

If it is possible to apply organizational structures to data in the way that Bayesian nets do, is the

inverse also a possibility? That is, can a model use observed data to induct organizational structure, and then apply it to a new set of the same kind of data? This kind of two-way relationship between observed evidence and structural organization intuitively seems very similar to what human learners do, and clearly much research supports the idea.

Kemp & Tenenbaum (2008) recently constructed a model that was able to use graphical representations of data – basically, Bayesian nets – to pick from a number of possible organizational structures the one that best fit the relationships in those data. They applied the model with apparent success to data as diverse as the hierarchical dynamics of a troop of mangabeys, the interactions between members of George W. Bush's first term cabinet, and trade relations between communities in New Guinea. In each case, the model chose an organization strategy that made sense for organizing the relationships in the data provided.

This work is still in its infancy, but already it suggests a future of grammatical modeling in which a sophisticated program is equipped not with pre-programmed structures for different knowledge domains, but instead with the ability to derive form and structure from ambiguous and chaotic data, which it can then use to aid in further learning. This is essentially a recursive loop, a system of feedback wherein learning can scaffold upon itself: its potential link to connectionist models based on exactly these principles is obvious. Today, we have only the earliest structure-deriving models, but as more sophisticated algorithms based on these principles develop, we may find we've unlocked a major door for artificial learning.

Connectionism, Math, and the Future

Obviously, the development of mathematics appropriate to systemic behaviors like recursion and feedback is a great contribution to any effort to model real, intelligent systems that depend so heavily on such behaviors. To the extent that we don't understand the math underlying cognition, or more basically underlying the biology and behavior of the brain that underlies cognition, how can we hope to model that cognition or biology, respectively? Identifying these algorithms is part and parcel of developing strong AI. Bayesian models offer some exciting future directions, but what new math will we uncover in years to come?

Presently, the partnership between mathematics, neuroscience, and computing is constantly strengthening. With projects like Blue Brain, we will soon have the computational power to at least test new theories and algorithms on a massively realistic scale, and with better brain imaging technologies and advances in neurobiological theory, we have a better idea every year of what goes on under the skull. One day, we may be able to work a neurodevelopmental component into these models as well – an evolving model of a brain over time, based on our increasing understanding of developmental neurobiology – reminiscent of COG with its distinct infantile learning stage. Where connectionist architectures and innovative mathematical models meet, the possibilities are impressive.

It is precisely due to this dependence on an understanding of cognition's mathematical

underpinnings that makes AI a joint endeavor between science and engineering, a task not only for clever programmers but for theoretical scientists as well. This is precisely what Herbert Simon meant when he said "We can stop debating whether AI is science or engineering; it is both." (Simon, 1995).

Analogical and Statistical DDAI

Composition as Computation

Emmy composes music. I've listened to several of her pieces, which can be streamed for free at www.artsites.ucsc.edu/faculty/cope/mp3page.htm. I would listen again, with pleasure. She is highly versatile; she can write in almost any style, given some time to familiarize herself with the material first. It seems that Emmy is a very creative musicologist. Unfortunately, her future as a touring musician may not be as promising as one might hope, since Emmy is a software program.

Emmy, or more technically EMI (Experiments in Musical Intelligence), is an AI program developed over the last few decades by composer and AI scholar David Cope. EMI may simply be a piece of software, but its product, its music, asks with its own kind of voice to be taken seriously. Hofstadter, himself a major proponent of analogical approaches to AI, expressed his own awe: "I was truly shaken. How could emotional music be coming out of a program that had never heard a note, never lived a moment of life, never had any emotions whatsoever?" (Hofstadter, 2002).

To the contemporary cynic of artificial intelligence, for whom every achievement of AI is only clever programming, this may not seem like anything very special. Through an historic lens, however, it is a remarkable accomplishment. The thing is, *EMI has musical intelligence*, at least insofar as one subscribes to the philosophy of the Turing Test – or at least a somewhat modernized version of the Turing Test. That is, that if a human judge cannot tell the difference between a machine's and a human's ability to exhibit intelligent behavior in a given domain, then that machine is therefore intelligent in that domain. In the same vein, if EMI can create sound which humans receive as music, and if music is a creative domain, then EMI is not only musically intelligent, but *creative by definition*.

Of course, a variety of objections can be raised to this perspective. EMI doesn't *experience* music, doesn't *think about* music. EMI doesn't even know the "rules" of composition in the way that Deep Blue knew the rules of chess, or in the way that many of EMI's predecessors in machine composition did. Cope doesn't feed EMI rules or techniques; he only feeds EMI notational data on a number of compositions in a given style, and from them, EMI analyzes the statistical behavior of the style and then composes new pieces based on those statistics (Cope, 1989). However, the pieces the program outputs are undeniably musical, and genuinely and emotionally evocative art, albeit with the help of the musicians who bring the music to life, and too with the help of the listener who brings their own internal world to bear.

DDAI and Analogical Theory

DDAI, or Data-Driven Artificial Intelligence, came into existence upon the meeting of two very different endeavors. One, signal processing theory, was originally only intended to create algorithms to sort out meaning from noise on high-static channels; I will discuss it in more detail in the section following this one. The other, analogical theory, set out to show that the process of analogical mapping played a key role in human thought. Their eventual combination is what led to the possibility of programs like EMI.

Analogical theory, which has emerged in the last few decades as a movement in cognitive research, extends from the idea that one element of high-level cognition is that people fit information into organizational structures. One of the leading theories of analogy is structure mapping theory (Gentner, 1983), which proposes that a major mechanism of learning and thought is the mapping between a source concept and a target concept: that is, the application of the known traits of and relationships between elements of one thing to make inferences about the traits of and relationships between elements of another. Ideas about the role of analogy in thought are at least as old as Greek philosophy (Shelley, 2003), but with Gentner and other analogical theorists, they've made a recent resurgence.

Hofstadter went so far as to identify the analogical process as "the core of cognition" (Hofstadter, 2001). In his view, analogy isn't only something humans use in high-level cognition; it plays a role even in basic perception. By applying mapped analogical structures, we actually encode only the symbolic representations from visual stimuli that we see – we only notice what we have learned is relevant (Chalmers, French, & Hofstadter, 1991). What humans end up perceiving isn't the same as what they see – a discrepancy that is, for example, the basis of illusion.

When analogical thought meets artificial intelligence, the result is the computational, statistical modeling of analogy. After all, humans can, even from early infancy, distill the core structural patterns within data using statistical co-occurrence and can generalize the patterns they discern to other data (Saffran, Aslin, & Newport, 1996; Marcus et al, 1999). This fusion of pattern analysis with analogical theory can be seen as the theoretical basis of DDAI.

As such, DDAI can be seen as a mid-level approach to modeling intelligence. It certainly is not as bottom-up and emergentist as the biological or neurological approaches of nouvelle or connectionist schools of thought, respectively. DDAI has very little to do with brain or biology, operating instead as a model of cognition or behavior. However, the approach is still substantially more bottom-up than most logic-based programs, which generally rely on a great number of pre-programmed rules as well as substantial after-the-fact corrections to “buggy” behavior. DDAI aims for about the same *level* of modeling as such rule-based programs, but does so with a single predominant, “neat” mechanism: a unified engine of statistical learning and pattern analysis, rather than a larger set of preprogrammed rules of operation.

Viterbi's Algorithm and Signal Processing

Since electronic communication media first became a reality, experts in signal processing have faced the problem of noisy signals. Over audio channels, when a signal becomes too corrupted, it becomes impossible to distinguish the message from the channel static.

Andrew Viterbi was a scholar of electrical engineering who was born in WWII Italy and fled to the USA with his Jewish refugee parents. In 1967, he received scientific renown for his signal processing algorithm, which helped to parse content from noise in audio channels. In Viterbi's own language, the algorithm *finds the most likely sequence of hidden events* – that is, the intended content of the message – given the sequence of observed events – that is, the actual electronic signals that come through to the receiver (Viterbi, 1967). The algorithm basically accomplishes this by breaking a continuous signal down into discrete chunks of sound, and by determining what "hidden", intended, audio signal each occurrence contains. That is, it breaks down a continuous signal into discrete pieces, and determines to what symbolic, theoretical state each is likely to correspond.

This process can seem complex, so I will put it another way. Viterbi's algorithm finds the core patterns within the ambiguous information, the method within the madness. With those patterns, it interprets – or, sort of, perceives – the confusing data. Does this sound familiar? In fact, the Probabilistic Graphical Models from Bayesian modeling, those same models which derive organizational structure from data, perform their inferences by using the Junction Tree algorithm, which is based on Viterbi's work (Jordan, 2004).

Viterbi's reach is far, indeed. His algorithm has since been applied to speech-to-text software – where it finds the phonemes or words that are hidden under continuous speech – to machine translation – where it finds the phrases in the target language that are hidden under the continuous stream of source language – and even to biochemical tasks like gene-finding that baffled traditional mathematical models – where it isolates the specific genes that are hidden in the continuous chain of genomic information (Rabiner, 1989 for speech-to-text; Brown et al, 1993 for machine translation; and Krogh, Mian, & Haussler, 1994 for gene-finding). A major part of adapting the algorithm to these tasks involved the collection of large corps of relevant data, in order to train the program, which is necessary so that the algorithm can establish the parameters by which it will break the continuous input into chunks and find the hidden events that correspond to those pieces.

Viterbi's Algorithm, and a handful of other statistical optimization techniques like Support Vector Machines and Kernel Methods, have turned out to have diverse applications considering how very simple they are compared to other approaches to AI. These systems do, essentially, one thing: they analyze data statistically to find their organizational patterns, and they make predictions based on the patterns that they find.

Modern DDAI

The basic operation of Viterbi's algorithm hasn't changed in the fifty years since it first enabled clearer translation of some fuzzy signals. The ways in which the systems have gotten better is mostly due, instead, to a combination of increased computational power and the massive digitization that occurred as information networks began to spring up around the world. Access to nearly infinite data opened up a new possibility for computational systems. By observing how humans, or other data points, behave in millions of instances, a computational system is able to better infer what behavior is salient.

Consider that a compositional program like EMI can analyze the basic style of classical composers by looking over a relatively small of pieces by Bach and Mozart: How then might its performance improve if it were also able to work out a general statistical understanding of musical theory based on the entire musical resources of the world wide web? Certainly, it is these massive, publicly available training grounds that have enabled the advancing performance of programs that can recognize handwriting, categorize texts and images, and make on-topic product recommendations in online stores (Agrawal, Imielinski, & Swami, 1993).

DDAI have become more sophisticated in another way: they have become recursive. Modern commercial DDAI, situated as they are within a dynamic digital marketplace, actually form recursive loops with their environments. They not only statistically analyze their environments, but they also act upon those environments.

For example, Amazon's product recommendation program reacts to consumer preferences by creating recommendations, but it also effectively acts *upon* the consumers when they take those recommendations. This creates a feedback loop, wherein the DDAI and the environment are engaged in an adaptive and continuous interaction and in which the system is storing information from past decisions as well as from present circumstances (Christianini, 2010).

A powerful, continuously operating DDAI, then, behaves much like a bacterial colony; it continuously stores information from its feedback loop within a dynamic environment and uses that information to inform decisions it makes towards a goal. Of course, the goals of a product recommendation program – making favorable recommendations – and a bacterial colony – survival – are very different, as are the environments from which they gather data, but the principles of agency are very much at work in both. In that sense, powerful DDAI placed within a complex, recursive system may represent a major component of future efforts in strong AI.

Furthermore, the most advanced contemporary DDAI are actually made up of sets of modular agents interacting together. Sophisticated e-commerce websites consist not of one agent performing its unitary goal within its environment, but of a set of modular agents interacting in the pursuit of related goals. For a website like Amazon or Google, a set of DDAIs work together, analyzing different sets of data and communicating with each other. These systems are engaged in dynamic feedback not only with their environment, but with each other (Christianini 2010) – an AI in charge of product recommendations may transfer information to and from an AI which analyzes website visiting

statistics and user demographics, and they both may alter behavior based on their interaction.

Such a modular AI harnesses systemic forces of recursion from within its own infrastructure – between the components that make it up – as well as from interaction with its complex digital environment. In the same manner that the differentiated colony of early jellyfish surpassed its single cellular ancestors in complexity, this class of multiple-component AI may represent the most sophisticated artificial system to date.

The IBM Watson, for example, was developed over the first decade of the 21st century and managed to defeat, in a widely televised match, the two world champions of the Jeopardy trivia game show. Perhaps the most striking event in AI since Kasparov's defeat, Watson's accomplishment was a far greater milestone than Deep Blue's which involved natural language processing, information retrieval, knowledge representation, and reasoning (DeepQA Project, 2011).

Rather than building a massive internal database of formal knowledge – like the knowledge-intensive Cyc program – or depending on a set of concrete logical rules as per the Problem-Solving Paradigm, Watson instead responds to every query by analyzing massive amounts of as-is information from existing encyclopedias and texts to generate many working hypotheses of varying statistical strengths. Watson analyzes keywords from the question as well as the clues from its syntactic structure, and compares this information to all the information at its disposal – for example, the entire text of Wikipedia (How Watson Works). Watson performs this task not once, but thousands of times in parallel using hundreds of different algorithms and strategies at once. Each search comes to an autonomous result, and Watson's confidence in its final answer is based on how many of its parallel searches ended up in agreement.

It's through this statistically fuzzy analysis that Watson handles the ambiguity of natural language and real-world information, and handles it far more successfully than any system to date. A set of interacting statistical components, then, in combination with the massively digitized interconnected information of our age, has been able to deliver results that, to date, no other approach has.

Other Examples

Advances in statistical optimization programs have led to search engines, language translators, library systems, statistical software programs, handwriting recognition programs, text categorization programs, computer vision, recommendations in online shops, genome analysis programs, and much more. None of these programs are meant to model strong AI: instead, they perform various useful tasks, in some cases with far greater facility and power than humans can.

However, not all DDAI are limited to traditionally robotic domains. EMI isn't the only analogical AI that is designed to mimic what amounts to creativity, and Watson isn't the only computer that baffles with its apparently humanoid behavior.

Aaron, for example, has been around for awhile. He is a robot designed by computer artist Harold

Cohen. He paints, using roughly the same kinds of combinationist principles that EMI uses to compose. Aaron operates as a DDAI that statistically analyzes graphical relationships based on works that it has scanned, and that outputs analogous art with robotic painting tools. Like EMI, though, Aaron's colorful, primitivist style has impressed many viewers, and a number of his pieces are on display at museums around the world (McCorduck, 1991).

Chatterbots are software programs designed to have conversations with humans, typically in a written text medium. These kinds of programs have been around for many years, and because the way they obviously and immediately test natural language performance, they've been very popular among the public. In recent years, with advances in DDAI, chatterbots have improved a great deal, but as early as 1966, researchers were designing software that could mimic human written conversation.

ELIZA, one of the earliest examples of artificial natural language processing, conversed with a very small knowledge base by relying on vague statements and questions. ELIZA's illusion was aided by the fact that the program was designed to simulate – or more accurately, parody – a non-directional psychotherapist at work. In such a role, ELIZA could often get away with asking vague questions and changing the subject abruptly, without breaking the façade of the role (Weizenbaum, 1966). Ultimately, the project probably taught us more about non-directional psychotherapy than it did about AI.

Eliza's successes – and shortcomings – led to the coining of the term "The ELIZA effect". ELIZA's namesake, the fictional Eliza Doolittle from George Bernard Shaw's play Pymalion, was a working class woman who passed herself off as upper-class by adopting an accent above her station. The aptly named ELIZA effect is the phenomenon in which people – including, with unfortunate frequency, project designers themselves – assign a level of internal insight or complexity to a system that is impossible given its actual programming and capabilities. This is especially prevalent in reactions to and claims about DDAI, which can achieve remarkable ambitions by piggybacking on statistical data from human beings. In that sense, one common objection to DDAI programs like Watson is that they perform so well for the same reason that too many of my classmates as undergraduates were able to pass their classes: that is, they cheat off their human neighbors.

One great example of the ELIZA effect in action is Cleverbot. Cleverbot was a chatterbot first developed by Rollo Carpenter in 1988. Its language-learning technique is very simple: build a historical database of conversational transcripts from its own experiences, and using a simple statistical analysis, respond the same way that humans responded given similar keywords. Carpenter and his team brought Cleverbot to the internet in 1997, and since then he has been chatting with any internet user who goes to his website, www.cleverbot.com. His conversational database has expanded from several thousand transcripts to over 20 million (Abumrad, 2011).

Cleverbot is an interesting program, and its limitations are as instructive as its successes. It can keep human conversationalists engaged for hours, but it can't respond with consistent appropriateness because it lacks both insight and the ability to use language creatively. Watson, the DeepQA system

that achieved success on Jeopardy, has surpassed all other DDAI in this regard, in part by using many more strategies on much more powerful hardware. Still, Cleverbot is a demonstration of how even a simple DDAI with a basic statistical processing engine and relatively humble computational resources can excel, given a large enough data base to learn from. It seems, then, that statistical regurgitation is a part of natural language, but not all of it; Cleverbot intrigues both because of the ways that it talks like us, and the ways that it does not.

Implications for Strong AI

In a way, DDAI was a historical accident. After all, Viterbi certainly never imagined the scope to which his algorithm would be expanded, and most researchers pushing the DDAI approach have generally ignored more theoretical psychological concepts like Gentner's structure mapping theory. Upon inspection, it seems that philosophies of analogy have created a theoretical basis for DDAI, even though most of the engineers actually creating DDAI programs with statistical tools haven't realized it.

From this perspective, the fact that statistical pattern analysis techniques applied so readily to problems of AI may not have been a historical accident at all. Statistical contiguity of occurrence is the key mechanism at work in conditioning and learning theory (Pavlov, 1927), and is well-established as highly present in learning scenarios for humans as well as other organisms (Skinner, 1938; Hull, 1943).

Statistical learning in the natural world isn't necessarily simple, either; covariation – that is, the predictive power of one event over another – is even more important than plain contiguity of occurrence in some learning scenarios (Rescorla, 1968). To only slightly transform Hofstadter's assertion about the centrality of analogy as the seat of cognition: it is possible that these techniques applied so well to AI because the statistical analysis of patterns – that is, the derivation of analogical structure – *does* play such a central role; perhaps it is, in fact, at the very core of cognition.

With the recent attention that Watson has brought to statistical approaches, and with the ever-increasing interconnectedness and digitization of information in the form of networks like the internet, as well as with advances in mathematical techniques like Bayesian modeling, the field of DDAI is looking at a lot of expansion in years to come.

What's less certain is whether analogical theory will keep up. To date, research in DDAI has largely been pursued because DDAI works; few DDAI researchers seem to have a sense of the relationship between their core techniques and the nature of the human cognitive engine. It seems to me that analogical theorists have asked many of the questions that get at exactly this relationship, but at this point, statistical theory and analogical theory have barely begun the kinds of conversations that both need from each other. Future collaborations between DDAI researchers on the one hand and psychological analogy theorists on the other might hold some very fruitful possibilities.

That said, there's a great deal of promise for DDAI as it moves into the future. If statistical-analogical progress continues at its current pace, it may soon become the predominant paradigm in AI,

at least for any models that emphasize knowledge acquisition.

Nouvelle AI

Shakey – What Robots Taught Us

Every AI approach discussed so far has been based essentially on modeling what happens in the mind and the brain, or at least on modeling behaviors very directly based on neurobiology and cognition. Nouvelle AI, on the other hand, is a collection of some very different ways of viewing artificial agency.

Early in the field's conceptual evolution, AI theorists from the Problem-Solving Paradigm assumed that logicism would extend naturally from artificial environments like the chessboard out to real, physical environments. Shakey, for example, was developed in the late 60's and early 70's at the Stanford Research Institute. It was the first successful analytic, goal-oriented robot, combining problem-solving capabilities with physical locomotion and the ability to sense its environment. In that sense, it combined many of the Problem-Solving Paradigm assumptions that dominated at the time with an embodied frame: a physical body. Given a command like "turn on the light", Shakey could look around, identify the light in question, locate a ramp to reach the light switch, push the ramp over to the wall with the light switch on it, and roll up the ramp to reach the switch. It did all this by searching through the problem space in roughly the same way that chess AI searches for chess moves (Nilsson, 1984).

Shakey generally impressed the public. However, over the next few years Shakey and other logic - based robots experienced enough limitations and delays in their capabilities that critics began to raise questions. As robots, they were "embodied", that is, they were situated within physical shells in actual physical environments, and the ambiguities of real perception and movement proved difficult for purely logical machines to master. In fact, part of the reason Shakey's designers were able to succeed at all was by crafting a stark operational laboratory environment for the machine that cut down on confusing cues as much as possible; in a more "real" world, Shakey would have been practically useless.

Unlike earlier logic machines, an embodied agent couldn't simply operate mathematical formula to move a pawn to D5. In order to act in its environment, a robot must – for example – discern the presence and the shape of a ramp using visual cues and distinguish it, categorically, from the appearance of a wall or a switch or a light fixture. Even with functional perception, a robot still has the hurdle of mobility to overcome – Shakey could only roll around on a smooth surface and operate a simple claw, because in its time no robots were yet capable of navigating the ambiguities of bipedal balancing or walking, especially given uneven or sloped surfaces, and complex robotic motor control was still many years off. Logical tools are based in definite, categorical rules: ill-suited for dealing with ambiguity. Agents like Shakey were very limited in their capabilities, and for a time, robotic approaches stalled.

In fact, it was only when researchers began breaking away from the classical symbolic approaches that further breakthroughs began to emerge in embodied AI. In the mid to late 1980's, roboticist Rodney Brooks produced a series of improving designs that quickly surpassed its logic-based competitors in locomotive abilities: a series for which operations were not based on symbols at all (Brooks, 1989). Brooks, and a growing and diversifying movement of others, contended that a great many of the tasks of life – like perception, motion, and balance – had nothing to do with symbolic, abstract rules, but instead came from the logically fuzzy, emergent, statistical behavior of interconnected networks. These revelations didn't only demonstrate the limitations of logic-based approaches, but they also demonstrated the richness of insights that the challenges of embodiment might bring.

In this way, Shakey, and early robots like it, began the journey away from disembodied, artificial, "floating minds", and towards a more nuanced understanding of system behavior within real, complex environments.

Nouvelle AI, as it has come to be called, developed over time from a growing, ubiquitous dissatisfaction with logicism's predominance in AI. Today, Nouvelle AI a rich and diverse collection of movement rather than a single school of thought, and is truthfully more easily characterized by what it isn't than by what it is.

Recall that Classic AI viewed intelligence as individual, rational, abstract, and detached. In other words, it assumed that cognition was housed within a single, solitary mind that was based entirely on the mathematical transformation of symbols, and that did not require a physical body capable of perceiving or acting upon its environment. This concept, the "floating mind", lies at the heart of Classic AI as the implied and generally unstated general model of intelligence. Nouvelle AI, beginning with the rise of robotics, would ultimately challenge each of these assumptions with alternative models of one kind or another.

Alternative Model One: The Embodied Mind

Many AI researchers had been frustrated for years by the vivid contrast between their successes in programming complex, logical tasks and their failures to model even simple sensorimotor performance. It was partly in response to Moravec's Paradox, as the contrast became known, that Nouvelle AI arose. Famed roboticist Rodney Brooks argued that high-level behaviors would be very simple for a system that had mastered the fundamentals of "acting and reacting" (1991). After all, biologically evolving organisms took billions of years to develop effective sensorimotor performance, and only millions to develop symbolic operations once sensorimotor performance was in place: why shouldn't researchers follow the same pattern? (ibid).

With these insights, Brooks and many others in robotics and AI shifted to a new philosophy of design: one that emphasized intelligence as a class of behavior that emerged organically from a system of relationships within a real physical environment. For this school of thought, cognition builds on

itself, and is layered from the real, physical interactions of simpler pieces into more complex ones – a philosophy of mind that starkly contrasts the rule-based logicism of the Problem-Solving Paradigm.

Robotics provided a rich domain for testing and refining these ideas, because for an embodied robot, deficits in effectual environment sensing and locomotive hardware, and deficits in the algorithms that control those sensorimotor operations, are made extremely obvious.

In the last few decades, the robotics industry has exploded. Robotics research today plays a central role in contemporary construction, aerospace engineering, and military technology, to name only a few. Better and cheaper hardware has become available, and robotic designs today are increasingly able to cope with real-world environments.

Funding has been especially available for research in applied robotics for the military. Big Dog, about the size and shape of a small donkey, is a quadrupedal robot by Boston Dynamics funded by the Defense Advanced Research Projects Agency as a near-future soldier's pack mule. The machine is agile, can traverse difficult terrain at 5 miles per hour, can recover its footing when forcefully kicked, and can carry up to 340 pounds of equipment (Railbert et al, 2008). Boston Dynamics is now working on Big Dog's successors, the Legged Squad Support System (Shachtman, 2008) – a larger, faster, smarter, and stronger Big Dog, the Cheetah – a faster quadruped agile enough to "chase and evade" (Rawnsley, 2011), and the Atlas, a robotic humanoid biped (Rawnsley, ibid.).

External prostheses are another area of development in robotics. Especially with improvements in biosensors – such that prosthesis users can control artificial limbs almost as smoothly as organic ones – artificial prostheses are reaching the point where their performance can match and even overtake organic limbs. Among the more recent advances are powered exoskeleton suits that can allow the disabled to walk or that enhance strength by many times – for example, to aid industrial workers. The Hybrid Assistive Limb (HAL) is such an exoskeleton, developed by Cyberdyne and Tsukuba University in Japan. HAL is a wearable robot that noninvasively senses the nerve signals beneath the skin and locomotes in tandem with its wearer's muscles. HAL and similar designs can help many disabled or weakened people to walk again, and are already available for testing at some care facilities in Japan (Robot Suit HAL, 2011).

When the resources for embodiment are not available, AI researchers have taken to sometimes simulating robotic bodies in order to get at some of the research advantages that embodiment offers. Simulated robots, or softbots, are AI programs that exist within a simulated physical environment. To use the technical terms, they are *situated* within an artificial environment to which they must adapt, but they are not *embodied* within an actual physical shell. SHRDLU, for example, was a contemporary of Shakey from MIT that was situated within a virtual – and manipulable – maze. SHRDLU was able to follow commands, moving about and transforming its maze environment, and was able to answer some questions typed to it in English (Winograd, 1971). In a lot of ways, softbots may seem to miss the point of embodied robotics altogether, but they are producible for a fraction of the cost of actual robots, and are sometimes able to simulate things that may be impossible given engineering and

budgetary constraints.

Softbots can teach us something about embodied behavior at least so far as their situated bodies and environments are accurate and detailed models. Because of this particular caveat, softbots are most useful for modeling simple robots, for which the lack of detailed physical laws and sensorimotor apparatuses does less harm. Research on large swarms of simple robots, for example, frequently uses softbots to test algorithms. Even in ideal circumstances, though, softbots have their issues as modeling tools. As the designers of the kilobots noted, "When using a simulation to validate an algorithm for a collective of robots, it is difficult to accurately model robots' interaction with each other, such as communication and sensing, and with the environment, such as movement and collisions. This modeling difficulty can lead to disparities in algorithm behavior when operating on a simulated collective versus a real robotic collective. " (Rubenstein, Hoff, & Nagpal, 2011).

Alternative Model Two: Nonhuman Intelligences

Close observation of animal behavior has shed light on human psychology for a long time, but it wasn't until the 1980's that hive insects began to emerge as a common model in AI. In 1991, in the same publication wherein he advocated for an emphasis on sensorimotor fundamentals, Brooks further made the suggestion that insects might be the ideal model from which to learn the basics of sensorimotor intelligence; from this idea, swarm robotics was born.

The key feature of natural swarms is that they manage to engage in intelligent behavior without a central guiding intelligence. In order to understand this school of thought, recall that from the beginning of this review, intelligence has been defined as broader than an anthropocentric notion of symbolic representation: intelligent agents are any systems that perceive their surroundings and act to maximize their chances of achieving their objectives. Each unit within a swarm is operating under simple algorithms that, in isolation, would result in simple behavior.

However, in synthesis with hundreds or thousands of other units, these simple algorithms result in much more complex behavior: a systems phenomenon commonly styled *emergence*, that unexpected, complex phenomena can emerge from systems of relatively simple, interconnected nodes. Emergentism plays a strong role in Nouvelle AI, especially in Swarm approaches.

Swarm Robotics, then, is concerned with how a swarm of relatively simple physically embodied agents can be constructed to collectively accomplish tasks that are beyond the capabilities of a single one. In the words of emergentist computer scientist Stephen Wolfram, "It's possible to make things of great complexity out of things that are very simple. There is no conservation of simplicity." Even the most complex life, after all, is layered upward from the interactions of simple molecular components.

Swarm research offers the possibility of understanding how these kinds of simple components interact to create complexity, an understanding that could open many doors for AI. In fact, research on the applications of swarm behavior to AI and especially robotics is a thriving subfield, complete

with current reviews emerging out of the international workshops of swarm robotics that began in 2004 as part of Science Advisory Board conferences. Erol Sahin and William M. Spears have published surveys of contemporary research on the topics, from the international workshops on Swarm Robotics that began in 2004 as part of Science Advisory Board conferences (e.g. Sahin, Spears, & Winfield, 2006).

More recently, the Self-Organizing Systems Research Group is a swarm robotics think-tank at Harvard University, and is responsible most currently for the Kilobot, a dream come true for budget-constrained swarm researchers. In order to make physically embodied research on swarm behavior accessible to more investigators, the kilobot was designed to be produced with only $14 worth of materials per unit – at least 10 times less than the next lowest-cost alternative – and to take about 5 minutes to put together from component parts. In addition, in order to overcome battery life issues and the unwieldiness of manually controlling hundreds of individual robots, the kilobot swarm works via a single overhead infrared transmitter that can operate on many bots simultaneously, and each kilobot has a 3-10 hour battery life – in batteries that can be charged collectively and without removal from the kilobots. These features make the kilobot a new breed of swarm robot: one that may allow researchers to experiment with large numbers of physical swarm robots, instead of necessarily choosing between using a small handful of swarm robots or a digital simulation only.

Each kilobot can locomote independently, can sense and communicate with its neighboring kilobots, and has on-board computation that can run any simple algorithm program sent to it by the central computer – a simple set of capabilities, but enough to test many phenomena in a new way. For example, kilobots have modeled algorithms like "follow the leader", where the robots travel in a long line, “dispersion”, where the robots spread out more or less evenly across a wide area, and even “foraging”, where the robots gather imaginary food and return with it to an imaginary nest, and much more (Rubenstein, Hoff, & Nagpal, 2011).

Of course, kilobots are not the only swarm robots entering the scene. Other areas of inquiry include self-assembly (e.g. Arbuckle & Requicha, 2010) and collective construction (e.g. Everist et al, 2004). Automated exploration is another vital area of research in swarm robotics (e.g. Parker & Sukhatme, 2006), with applications for unmanned missions in environments hostile to human life. Lingodroids, for example, author an aural language as part of their swarm algorithms, coming up with shared names for locations and directional instructions as they use their social interaction to collectively explore and map an area (Schulz et al, 2011).

Unlike lingodroids, however, most approaches to swarm robotics rely in no way on symbolic representation. When Newell and Simon authored the Physical Symbol System Hypothesis, as reviewed previously, they posited that “A physical symbol system has the necessary and sufficient means for general intelligent action.” One of the key differences between a robot like Shakey or a Lingodroid, which are both based more or less in logicism, and a swarmbot collective like a hive of kilobots is the presence – or lack thereof – of symbolic representation. Shakey navigated its surroundings by holding, in its "mind", a map of its environment. Like a sailor or a pilot, Shakey kept

track of landmarks and plotted where it was with a "You are Here" flag on its internal map. A kilobot, however, holds no internal representation: it operates with simple rules based on its operational algorithms and the things that are present in its immediate environment. If a kilobot can't sense an object in a given moment, then for the kilobot, that object doesn't exist – there is no referential map, no storage of symbolic memory.

The striking thing is that even without symbolic representation, a swarm of kilobots can surmount the same kinds of challenges as robot with representation, and in many cases can do so more successfully than the logic-based bots. Rodney Brooks explained the striking successes of non-symbolic robotics succinctly: “The world is its own best model. It is always exactly up-to-date. It always has every detail there is to be known. The trick is to sense it appropriately and often enough.” (Brooks, 1990).

The argument here is not that symbolic representation doesn't play a role in higher cognition, but that it isn't the only mechanism at work. Proponents of non-symbolic approaches to strong AI argue that a great deal of research has been invested in the algorithms of logic and symbols, and that there is a much more than logic at work in intelligent systems. Non-symbolic algorithms don’t only play a role in swarming organisms, after all: they are at work throughout our own cellular and nervous behavior.

In their own review on the subject, Bonabeau, Dorigo, and Theraulaz (1999) articulated the hope that a distributed approach like swarm robotics holds for many in the field of AI. “At a time when the world is becoming so complex that no single human being can really understand it, when information (and not the lack of it) is threatening our lives, when software systems become so intractable that they can no longer be controlled, perhaps the scientific and engineering world will be more willing to consider another way of designing “intelligent” systems, where autonomy, emergence, and distributed functioning replace control, preprogramming, and centralization.”

The movement towards swarm algorithms, rather than centralized processing, mirrors not only neurobiological process, but biological evolution as well. As Christianini (2010) discussed, “How do we combine large numbers of noisy (imperfect) modules, and drive them with noisy data, and still obtain reliable behavior? The answer is in properties of the system, not of the individual modules. Natural systems must have faced and solved this problem too.” In other words, human intelligence derives fundamentally not from private rules but from interactive, systemic effects, both within an individual and between an individual and their world.

Complementing the progress made by roboticists is the ongoing contributions of biologists studying swarm behavior. By studying ants, bacterial colonies, and many other examples of swarms in the natural world, biologists are uncovering more and more of the algorithms that evolution has created to solve the problems of survival and propagation.

Many swarm robotics researchers hope that over time, the emergence of complex behavior from swarm algorithms will scaffold upon itself. That is, that the identification of the algorithms at work in

simpler swarms will open doors to models of behavior in more sophisticated systems. After all, just because a system is built out of simple pieces, does not imply that it will manifest simplistic behaviors. In the words of Charles Darwin (1859),"There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved." In this way, nouvelle AI, more than any other domain of AI research, captures the bottom-up ethos as a rule.

Alternative Model Three: Developmental Environment

Another nouvelle perspective, one that is just now emerging in AI, is an approach to intelligence that looks for sapience in the longitudinal, relational process between an individual and its social and physical environment, rather than looking just in the unitary individual itself. It seems self-evident that the most sophisticated AI of the future will probably be those that are able to learn fluidly. The developmental approach to AI capitalizes on that concept by advocating for the necessity of a stage of real-world learning before a robot is fully functional.

For example, COG is a project at MIT that emphasizes social development as part of AI design (Brooks, 1998). Human intelligence, from the perspective of COG's designers, emerges not only out of one's brain, body, and physical environment, but out of one's social interactions with engaged fellow humans.

To model this, the researchers on the COG project have created developmental platforms analogous to child or infant forms of the "adult" COG. One of these, Kismet, is designed with two capabilities relatively novel in the robotics scene. First, it has the ability to use its robotic face to express a number of motivational states – that is, simulations of emotions. Second, it has the ability to make eye contact with its "caretakers", as they call themselves. Eye contact may seem like a minor detail, but in fact it seems to be a fundamental building block of intelligence in humans and nearly nonexistent in AI research.

Historically, complex social processing of any kind has been almost entirely ignored in AI research, an oversight that may have been a mistake. For example, eye contact is how developing humans manage to surmount the seemingly impossible joint-inference problem (Quine, 1960). In learning language, infants face the linguistic indeterminacy of not knowing which of the many sounds they encounter matter, and which of the many objects they encounter such sounds are referring to, a problem that has been puzzling for AI researchers as well as linguists. Toddlers, with only limited exposure to language, manage to overcome this in part by using social cues (Baldwin & Moses, 2001, for a review). For example, infants overcome potential linguistic ambiguities by following a speaker's gaze (e.g. Baldwin, 1993; Tomasello & Barton, 1994). Social sensitivities like gaze-monitoring make up some of the earliest traits identifiable in infants, and are probably to some extent innate (Baron-Cohen, 1995). Given social cues, an 18-month-old human will readily master a new word after a single trial

(Nelson & Bonvillian, 1973), a feat impossible without such cues.

Social skills may be just as essential as cross-situational statistical tracking in an infant's ability to learn advanced human language, and in fact are almost definitely used in tandem with such statistical detection. For example, the machine-translation model of language learning mentioned earlier in this review (Yu & Ballard, 2007) learns by using Bayesian inference to analyze both the usual cross-situational associations and by using social cues to infer speakers' intent, as do other cutting edge models (e.g. the intentional model for language learning designed by Frank, Goodman, and Tenenbaum, 2009). This type of design consistently outperforms word-learning models that neglect social cues (Frank, Goodman, and Tenenbaum, ibid.)

Such innate social sensitivities may represent one of those sets of algorithms that must be innately present in any system for which semantic mastery is a goal. Without social processing, even an otherwise fully-articulated general learning program will at best behave with extreme autism: showing no motivation or ability in the social elements of learning which, for humans, motivate a lion's share of development (Baron-Cohen, Baldwin, & Crowson, 1997). As current models become more and more complex, models of social learning should come to occupy a more central role in AI research, especially that research concerned with language. In the long run, swarm robotics – with its emphasis on intelligence through interaction – might lead to great breakthroughs in this area.

Social processing is also intimately connected with the development of complex emotions. For example, gaze-monitoring in a variety of situations is one place where infants can learn about appropriate emotional responses to their environments (Moses, Baldwin, Rosicky, & Tidball, 1999). Contrary to the flat-affect AI popularized in early science fiction and likely encouraged by the field's origins among a 1950s peer group of computer scientists who were probably more likely than most of us to see chess as the apex of intelligent behavior, it seems likely that complex socio-emotional development is another central pillar of articulated consciousness.

Alternative Model Four: Intelligence in the Spaces Between

Robotics challenged classical theories of AI by advocating for the role of the physical body in intelligence; swarm robotics took this a step further by placing intelligence not *within* a body, but between many of them. Transhumanism – literally, that which is beyond the human – is not so much a scientific approach as a collection of radical challenges to many of our basic assumptions about intelligence. It sheds further light on nouvelle views of AI by viewing intelligence as a phenomenon that emerges from the space between the constructed boundaries of discrete humans and their environments. In other words, artificial intelligence can be perceived as any intelligent behavior that's based, at least in some part, on technology. From this perspective, any technology that artificially augments or transforms the normal capabilities of biological agents also qualifies as AI technology. From a transhumanistic stance, then, human : machine interfacing is a subdomain of AI research.

Clark (2003) articulated the philosophical foundations of transhumanism extremely well with the

example of a wristwatch (page 41). If a man wearing a wristwatch is approached and asked "Do you know the time?" he is likely to answer in the affirmative before even checking his time-keeping device. He knows the time already, he claims implicitly, even though that knowledge isn't actually in his mind yet. If, as he looked to glance at his watch, one reached out and covered his wrist, he would be hard-pressed to explain why he had claimed to know the time a moment ago – it is the time-keeping device that "knows" the time, not the user, but modern culture humans have so completely integrated with the technology of the wristwatch that they consider something that the wristwatch "knows" a part of their own knowledge.

This is not true with all technologies – it is based on factors like speed of accessibility, ubiquity of use, and how long the technology has been around. Even if we carry a pocket dictionary everywhere, for example, we do not answer in the affirmative if someone asks us whether we know an unfamiliar word. But over time, humanity has the ability to integrate technologies such that we identify their capabilities as part of our own capabilities – if an external vocabulary device became as pervasive and rapid-access as the wristwatch, for example, might not most users be claiming knowledge of every word in our language in a century or two?

When we identify machines as a part of ourselves, when we have integrated them seamlessly into our functions, then we have become cyborgs: beings with both biological and artificial parts. In that sense, humans have been becoming cyborgs since the first tool was lifted and used, and we have had artificial intelligence – by this cybernetic concept of it – since the first paintings appeared on caves and augmented our biological ability to store and process information. Kurzweil (2008) extended the idea with what has become a piece of the transhumanistic vision of the future: "...our computers are not going to be these rectangular devices we put in our pocket. They`re going to be inside our bodies and brains and we are going to be a hybrid of biological and nonbiological intelligence."

For example, neurocognitive prostheses are an idea that emerged from science fiction as an application of the "cyborg" concept to the brain, rather than the body. With artificial aid, the concept's advocates have speculated, mental abilities could be enhanced as easily as physical ones, leading for example to better attentional abilities or increased memory capacities.

Until very recently, this idea was speculation only. This year, however, cutting edge research at USC identified what seems to be the key mechanism in the encoding of long-term memories in rats. With this information, the researchers were able to cleanly block that encoding via a targeted drug. Then, far more strikingly, the researchers created a neural prosthesis that could connect to the drugged rats and replace the function of the biological mechanism that they had blocked. With this, they could turn the rats' ability to remember on and off by turning the device on and off. Perhaps most exciting of all, when the neural prosthesis was installed in normally functioning rats, it actually improved their memory performance significantly above baseline.

Plans for the future in this line of research include testing of the memory expansion device on primates, and it may not be long before we see memory-enhancing prostheses available for humans.

The paper is under review for the Journal of Neural Engineering, and when published it will be entitled "A Cortical Neural Prosthesis for Restoring and Enhancing Memory." Authors include Sam Deadwyler, Theodore Berger, Vasilis Marmarelis, Dong Song, Robert Hampson, and Anushka Goonawardena.

Transhumanistic ideas have become the fodder of contemporary science fiction, with their rich and evocative possibilities of an imagined future. However, in the new millennium, distant-future speculation seems hardly necessary; the contemporary science of human : machine interfacing, which Clark (2003) reviews in some detail, is no less exciting. With smaller and smaller digital devices, *wearable computers* are not far off – these may take the form of a pair of sunglasses with the capabilities of a modern laptop. This type of computerized-eyeglasses technology could easily be paired with *augmented reality*, technology that overlays sensory details or information atop of one's perception of a real, physical environment. *Tangible user interfaces* represent another approach to pervasive technology, wherein daily objects are increasingly synthesized with computer technology. More and more sophisticated digital art tablets are one example of this; we aren't far from engineering a digital tablet that feels like and can behave in nearly every way like a high quality painting tool set, but without the need for actual paint or canvas.

Futurists and transhumanists tend to liberally synthesize AI research in the service of their visions; they are often highly theoretical and sometimes wildly speculative. For many transhumanists, the future represents a time when research in genetics, nanotechnology, and robotics – collectively GNR, as in the GNR Revolution – merges into a single unitary understanding which overcomes all the current limitations of humanity – death itself, for example – and brings us into a new golden age. Kurzweil himself, probably the most influential transhumanist who has yet lived, frequently relies on scientifically unstable grounds in his claims about the future and has subsequently become a discredited figure among the vast majority of empirical researchers. However, though Kurzweil's timeline and the details of his future are probably well off the mark, he has nevertheless created an impressive vision of the possible that is grounded in current technology and that offers some compelling ideas about the direction we are moving, and where it might take us.

Discussion

Without the Dream

When I introduced this review, I proposed the value of *utility*: that is, the usefulness of the strong AI dream.

In taking a closer look, it becomes starkly clear that strong AI has played a central role in the field's history. It doesn't take much imagination to realize that, if the dream of sentient machines had never been, the field of AI would be very different today. Most likely, without the unified vision, there could be no real unified *field*: only the various endeavors that I've reviewed so far, like robotics, mathematical modeling, computation, etc., but without the link of their relationship to a unified progress. The strong AI concept has made the AI vision comprehensible, to experts and the public alike; when the dream is missing, it handicaps public discourse and understanding of the field.

Because it facilitates public understanding, academic engagement with the dream of strong AI can actually protect from the downfalls of ungrounded promises about the timeline of progress. Furthermore, a great deal of funding for AI research and technologies comes from sponsors for whom the potential of strong AI systems is a compelling vision, and the unified goal that strong AI brings has also been good for creative collaboration, giving researchers and engineers a reason to work together and discouraging unnecessary sectarianism.

Without the dream of strong AI, these disparate subfields would have had little reason for theoretical communication, and even less common ground upon which to build working relationships. There's no doubt that some collaboration would still occur at their fringes, but on the whole, robotics would be robotics; swarm research would be swarm research; and psychological theory of cognition would hold little significant bearing on robots or computers, to name a few examples.

To really grasp the role that the dream has played in our history, let's perform an experiment with the kind of speculative fiction that has been so important for AI, and take a moment to imagine a world without the dream. Let's imagine a world in which supercomputers are only designed for the purposes of running large science and engineering projects, and developing faster hardware for gaming hobbyists – a world wherein Arthur C. Clarke doesn't lurk in the public awareness behind every large-scale computation project with his nightmarish yet exciting HAL gone haywire; in this world, nouvelle AI is only about robotics, only about semi-automated tools for military and industrial purposes; the emergent behaviors of a swarm are not the first stirrings of a collective intelligence, they are only bugs in the programmers' code; modeling approaches are only about developing and testing theories of various mechanisms of behavior – psychological metrics only, where no one dreams of breathing into them life.

For some, those for whom acute specialization is comfortable and those for whom a world transformed by technology is a fundamentally frightening thing, this arrangement may feel more suitable than a unified goal of strong AI. However, from a perspective that favors progress in science and knowledge, *the concept of strong AI has served a vital role even if its dreams are never to be realized.* Even if for some reason strong AI is a doomed project, the collaboration and theoretical synthesis that it has yielded has been invaluable to the advancement of general science. Robotics, computer science, and even the entire cognitive aspect of contemporary psychology, owe huge parts of their foundations to concepts from work inspired by strong AI. In considering the theories and projects from the previous review, it becomes plain just how many of them have depended upon collaboration or debate involving different schools of thought, brought together in order to navigate the nature of and the best route to strong AI. Much progress has been a result of the reactions and interactions that the strong AI concept has made possible.

Historically, then, strong AI has played a major role in the progress of the field. But what is the future of the dream?

In the introduction to this review, I discussed that the dream of strong AI today has weakened considerably, and has become an unpopular one to openly share. As the subfields of AI research have further specialized, this unification has weakened further. In another decade or two, without this dream at its center, will the conceptual frameworks that make up AI theory have fractured so much that it becomes impossible to author a theoretical synthesis like this one? Will these subfields continue specializing with diminishing collaboration, and without the common language that such collaboration engenders? Or will strong AI make a popular resurgence – will scientists and laypersons find themselves unashamedly excited by its possibilities, again?

If current trends continue, and the unifying dream does fade away, I contend that scientific endeavor as a whole will be the poorer for it. Without the dream, the timeline of strong AI's realization will likely be set back considerably, but the conceptual impoverishment to which I refer extends well beyond that specific project. The impoverishment to which I refer is the death of a rich and varied exchange of ideas. Without strong AI, there is no shared problem; each of the approaches herein becomes exclusively concerned with its own narrow goals: each a ship passing the others in the night. When their proponents argue, on the whole, it can only be a struggle for academic territory. After all, how can debate be fruitful when there is no shared task to debate about, and no shared language? Interdisciplinary collaboration and communication make synthesizing vision possible, and vision is the fertile soil so necessary for the nurturing of any profound, revolutionary work.

With The Dream

With strong AI at the center, the shared problem becomes apparent as the self-evident quest of the dream itself – the realization of strong AI – and every one of the approaches covered herein transforms conceptually from competing theories to become a unique and potentially vital angle for movement

towards that vision. Debate and discourse, then, even when spirited, become the discourse of fellow pioneers, in new territory, arguing heatedly over their shared maps.

Much earlier, I mentioned the concept of a General Agent Architecture: an artificial system that can synthesize many layers of various operational subsystems into a single global agent. Mostly fleshed out by robotics researchers at MIT – who developed a specific approach called Subsumption Architecture in order to incrementally add behavioral complexity to already-functional robots (Brooks & Connell, 1986) – the technique was later adapted by designers of sophisticated game AIs. However, the General Agent Architecture's core principle of modularity holds a rich promise for strong AI more generally.

Human cognition, after all, doesn't seem to be based on a single rule or technique, but the combined functioning of various distinct algorithms for memory, logical analysis, imagination, social processing, sensorimotor processing, and a great many more. The algorithms and subsystems that are well suited for symbolic logic, for example, can only very clumsily perform the distinct demands of sensorimotor tasks, as early failures in robotics aptly illustrated. Our sensorimotor algorithms, on the other hand, though superbly adapted for balancing on two feet, have no way to even begin puzzling out chess strategies. Within the human system, somehow, this diverse set of algorithms and skills is unified into a single overarching system: a general agent architecture.

It is compellingly easy to argue that any fully realized strong AI must follow such an architectural strategy. It is more difficult to comprehend in the particulars how such a system might actually fit together, especially given our relatively infantile understanding of how we humans work. Until a puzzle is complete, how can its picture be clear? Until such a system is a working reality, we can hold no definitive knowledge of its make-up.

However, it is possible to make some guesses, based on what we know today. Within every section above, I offered examples of both successes and limitations of each approach, and suggested some pieces of what those developments might be teaching us about the ultimate shape of strong AI. Within this final synthesis, I want to draw on the whole of the preceding review in order to tentatively place the pieces of each approach where they might fit within a theoretical, comprehensive general agent architecture of tomorrow. Of course, such a synthesis can only make sense from the perspective of the strong AI dream, in which each approach is seen clearly as working on different aspects of the same problem, different pieces of the same answer.

The Character of a Next-Generation General Agent Architecture

Upon review, perhaps the best way to envision the theoretical tasks of designing a General Agent Architecture is to divide them conceptually into two major categories. On the one hand, there is *what must be performed*. That is, what tasks are necessary for strong AI to be possible? What must the algorithms accomplish? On the other hand, there is *how it will be performed*. That is, what type of mechanisms will make those algorithms executable? What will enable their development and

enactment? To neither set of problems do we have definitive answers. However, the majority of work in AI so far can be conceptualized as efforts to move forward in one or both of these areas.

I'll begin with the first set: *what must be performed.* Much of the work mentioned in this review can be perceived as movement forward in one of the following types of tasks: logical thought, sensorimotor skill, analogical graphical representation, or statistical pattern analysis. More recently, socioemotional processing has begun to emerge as another discrete type of task essential for strong AI. Any task within the field that I can think of can be imagined as an amalgam of these skills. Emmy's musical compositions, for example, are a pure exercise in statistical pattern analysis. Shakey the robot used logical rules and early sensorimotor processing algorithms to try to move around. Bayesian language learning models or the IBM Watson, both more complex systems, use logical rules, pattern analysis, and analogical structures in interactive combination to accomplish their goals. These five core task types can be seen, perhaps, as the *engines* of cognition: they may not all be necessary, and they may not be sufficient, but at least they are a good starting place given the current state of knowledge in the field.

As the theoretical root of AI, the Problem-Solving Paradigm taught us a great deal. The successes of the approach were, of course, concentrated within the domain of the logical thought *engine*: but in fact, in the long run, the approach's failures in other areas taught us nearly as much as its successes. Today, our logic programs are very sophisticated, able to conduct mathematical operations and logical puzzles with complex and multilevel reasoning skills. Some researchers, the proponents of GOFAI, still believe that these types of logic systems will eventually – in themselves – yield strong AI. Within a General Agent Architecture for strong AI, logical thought has an undeniable role, and the successes of the Problem-Solving Paradigm means that we may already have all the necessary tools to design the logical components for such a system.

More recently, robotics researchers have pushed for better and more affordable hardware, and in combination with advances in brain-machine interfacing, we are truly seeing the beginning of a cybernetic age. With more and more complex robotic bodies and better design of the algorithmic *engines* of sensorimotor processing, we are closer every day to robots that are able to interact seamlessly with the real world. Most likely, if the lessons of robotics movement taught us anything, it is that a general agent architecture for strong AI will not only be the "floating brain" imagined by most researchers from the supercomputing and problem-solving approaches. Instead, it will involve discrete sensorimotor processing and some kind of physical embodiment with which it can exercise that processing in the real world. After all, in every actual working model of strong AI that we have – that is, the human model – the gaining of intelligence involves sensorimotor interactions as well as internal processes.

Until recently, cognitive approaches to AI largely fell within the domain of logic and problem-solving. However, two new ideas are emerging to supplement rules-based reasoning. The first, which falls under the *engine* of analogical graphical representation –as typified by Bayesian structural modeling – allows a hierarchical organization of ideas into structures. Hopefully, current Bayesian

techniques are only the beginning. As mathematics researchers begin to work more closely with psychological modelers and neurobiologists, who knows what kinds of structure-organizing techniques they will create in years to come?

The other major new cognitive approach, DDAI, often fits unobtrusively in support of other AI approaches. The *engine* of statistical pattern analysis is a relatively simple cognitive ability that may be central to enabling the acquisition of a great many more sophisticated skills, in natural as well as artificial systems. Paired with analogical graphical representation, for example, it allows the derivation of the core structures from datasets that contain a diverse many conceptual organizations.

Finally, recently, AI researchers have begun exploring the importance of a socioemotional processing *engine* in the emergence of AI capabilities. In Nouvelle AI especially, researchers are designing robots with the ability to track gaze and to simulate and express emotional cues within a social context. Given the central role these skills play for developing humans, it is likely that these advances have much to teach us about the process by which intelligence develops within a complex social environment. This research is only just beginning, but in time we can hope to learn more about the traits that are necessary within a social environment, as well as within a learning system, that make the development of a strong AI system possible.

The Development of a Next-Generation General Agent Architecture

The second main classification of AI work can be conceived as the *how it will be done* work. That is, what type of systemic mechanisms and research methodologies will make progress possible in the kinds of tasks discussed above? Upon review, current major efforts at improving methodologies or designing systems that can execute the tasks of AI seem to largely break down into four areas: supercomputational resources, connectionist simulations, emergentist systems, and transhumanistic technologies. Unlike the five thought engines laid out above, these methodologies in no way suggest a totality of available methodologies; we don't even know if they will end up being very good approaches. However, many are promising, and together they characterize most of the current work in the field.

The persistent and exponential increase of computational resources may be the most certain movement towards strong AI. Better and better hardware often plays an enabling role for theoretical advances in other AI approaches; luckily, so far, better and better hardware has been constantly available. Genetic algorithms, comprehensive neurobiological models, DDAI with massive databases, and others: all of these systems will especially benefit from accessing computational resources beyond those currently at our disposal. If Moore's Law continues in its effects, then it may only be two or three decades before *the average home computer* has the computational power to simulate the human brain. By that time, advances in quantum computing may begin to play a role in further miniaturization, and may accelerate the advent of computational processing power even more.

As reliant on computational advances as any other approach, connectionist modeling represents another route towards strong AI progress. Large parallel processing networks show some qualities that other forms of computation do not, and – especially in the case of neuromorphic nets – possess the distinct advantage of closely resembling neurobiological architecture. Out of such neuromorphic and recursive systems, unexpected properties can emerge.

This theme of emergence has become an increasingly popular one in other approaches to AI, as well. In Nouvelle AI, many researchers have argued that biological evolution can provide us with a powerful model for how intelligent systems may scaffold up from simple forms into more and more complex ones. Robotics research on swarms is based on the same principle: the emergence of complex phenomena from simpler pieces. Using the genetic algorithm approach, other researchers hope to harness mechanisms of selection and emergence to develop increasingly complex artificial systems. This bottom-up approach may eventually lead to an understanding of intelligence that reaches down to the molecular level processes at the root of intelligent agency.

An alternative manifestation of the theme of emergence turns up in *development as design* – a philosophy that has begun to play a role in robotics with projects like COG introducing immature, infantile, learning stages for their robots. Furthermore, as connectionists and modelers come to better understand the details of neurobiology and its development throughout childhood, we may eventually

be able to incorporate this neurodevelopmental information into computational models. After all, developmental differences at different ages certainly play a role in human development – a phenomenon which probably explains the critical periods of language acquisition and why victims of early extreme social isolation are later unable to become fully fluent language speakers (Curtiss, 1977). It seems, based on such evidence, that if knowledge-less newborns were equipped with neurologically adult brains from birth, sentience would probably be impossible. Perhaps one day, once we've learned a great deal more about the brain, it will be possible for us to create artificial agents for which a simulated developmental neurobiology develops in tandem with knowledge acquisition.

Finally, with advanced cybernetic prosthesis and the advent of the first external memory devices, the lines between the natural and the artificial blur more every day. If these lines blur enough, if scientists are able to create technological devices that can interface with many aspects of human neurobiology, then it's easy to see to see how such advances could be reverse-engineered and applied to strong AI. For example, if human memory can be stored on artificial digital prostheses, then could a future version of such a prosthesis be adapted such that its memory could accessed by an AI system? After all, AI may not only exist in a discrete cognitive system, but in the transhumanistic *space between* technology and natural systems.

With at least five different conceptual engines of cognition, and at least four major areas of broad methodological thought present in the field, there is no easy way to predict exactly how AI research will unfold in the near future. In the years to come, it will be fascinating to observe which of these conceptual frameworks bear fruit and in what ways. However the technology develops, and whatever is discovered, this research has the potential to teach us a great deal: not only about artificial intelligence, where having read this review the applications should be obvious, but also about the nature of intelligence itself. Insofar as the endeavor of AI ultimately represents a modern manifestation of this perennial quest of humankind - that is, our quest to understand, copy, and create ourselves - the unfolding of AI research over the next few decades may illuminate a great deal in psychology and the social sciences that has heretofore been unclear. A detailed simulation is an extremely powerful tool for understanding, so as science and computational resources begin to rise to the challenge of the inquiry, what will be discovered about the nature of a simulation of the mind? When we can answer that, I imagine, we will have a much stronger understanding of our own consciousness than we do today.

In Conclusion: The Top-Down and the Bottom-Up

Typically, when top-down and bottom-up theoretical orientations are mentioned in the same discussion, it is to compare their merits, competitively. Here, I set out instead to explore how, when unified by the idea of strong AI, top-down and bottom-up approaches actually support and complement each other.

Early on, I also mentioned that the distinction between top-down and bottom-up approaches was not as stark and clear-cut as it often seemed. Consider: from the framework of a General Agent Architecture for strong AI, the top-down goal can be summarized as the identification of tasks that must be performed by an agent at various levels of its operation. For example, by modeling the workings of language and grammar in humans, we can discover what kind of things an agent must already know in order for language to be learnable. Above, I mentioned logical thought, sensorimotor skill, analogical graphical representation, statistical pattern analysis, and socioemotional processing as potential *engines* that might be necessary parts for any successful model of strong AI, at least based on the current working knowledge of the field.

Over time, as obvious high-level tasks become more clearly differentiated and modeled, top-downers should be able to move further and further down toward more fundamental processes. For example, in the future, top-downers may have successfully identified the algorithms at work in complex movements like dance and balance, and may then be able to move down and identify exactly what roles different anatomical structures within the brain play in the execution of these same behaviors. In this way, top-downers can move towards modeling more basic processes over time.

In the meantime, the bottom-up goal has been stated as thus: to model as complex, intelligent behaviors as possible with as simple, basic algorithms and starting conditions as possible. For top-down approaches, constant pre- and post- programming – that is, constant *hacking* of the system – is a permissible way to get results. For bottom-uppers, adherence to the principles of complex systems as they evolve and behave in the natural world – that is, adherence to *high-fidelity* – is far more important. What has gone largely overlooked is that top-down work has the potential to greatly assist bottom-uppers in this task, by identifying target behaviors for which a bottom-up approach might aim. With their help, bottom-up researchers can reach for such behaviors or tasks from an emergentist toolset. For example, if the goal of bottom-up researchers is to model how a connectionist learning system can acquire and use a complex set of grammatical rules without pre-programming those rules into the system, then it becomes highly valuable for the researchers to know exactly what those rules are. That is exactly the kind of information that top-down researchers are equipped to discover and provide.

Even as top-down research moves down through layers of complexity, so the fruits of bottom-uppers' labors will become more and more sophisticated and high-level. Eventually, bottom-uppers will no longer be limited to modeling systems with the complexity equivalent to a small colony of ants,

using a swarm of ant-bots. Instead, they may be seeking to model the behavior of a mind, which is based on the collective behavior of simulated memory, analogy, pattern analysis, logic, etc, which are based in turn on the collective behavior of simulated areas of the brain, and so forth.

It becomes clear, then, that some day the top-down approach will meet the bottom-up approach; top-downers may be working to uncover the algorithms behind the neurobiology of language, for example, while bottom-uppers work to create a system that manifests creative language behavior. Over time, assuming that they can still understand each others' work, it will become exceptionally clear to each camp that the others have been performing work that they need to proceed. Top-downers will have been developing the knowledge and skills to point out an outline of research goals, of problems that need to be solved, while bottom-uppers will have been developing the knowledge and skills to design algorithms and agents that are both faithful to natural systems and highly effective at reaching those goals.

In other words, as these many various approaches work on the various disparate problems of strong AI, it is likely that they will eventually find themselves holding the keys to each others' breakthroughs. The mathematical techniques of Bayesian graphical modeling may end up as the next big piece in the DDAI puzzle. Swarm algorithms developed in Nouvelle AI laboratories may be exactly what's missing from the next generation of comprehensive neurobiological models. There is absolutely nothing wrong with the fact that the field has specialized and diversified: the plethora of rich ideas that I've presented here are a testament to that fact. But without a shared vision, without a common synthesis, how will these fields come together in collaboration, once again? Before the specialized domains split too far apart, let's remember the AI dreams and challenges that the field shares. Synthesis in such a widespread field is difficult, yes, but it is also possible, and it is from such synthesis, sometimes, that the best cross-disciplinary ideas can be born.

I began this synthesis with the words of Hans Moravec (1988), and I will end with the same. He said: "I am confident that this bottom-up route to artificial intelligence will one day meet the traditional top-down route more than half way, ready to provide the real world competence and commonsense knowledge that has been so frustratingly elusive in reasoning programs. Fully intelligent machines will result when the metaphorical golden spike is driven uniting the two efforts." The technology with which we can forge Moravec's golden spike is nearer to us every day, with progress from the bottom and from the top. In the meantime, the best hope for a vital field of artificial intelligence lies with researchers who embrace the vision of a cross-disciplinary effort towards strong AI: that is, researchers who understand the importance of a common language, a common goal, and a common dream.

References

Aaboe, A. (1974). Scientific astronomy in antiquity. *Philosophical Transactions of the Royal Society 276 (1257)*, 21–42.

Abumrad, J. (2011, May 31). Talking to machines. *Radiolab*. WNYC. Retrieved 07/04/2011 from http://radiolab.org/2011/may/31.

Agrawal, R., Imielinski, T., and Swami, A.N. (1993). Mining association rules between sets of items in large databases. *SIGMOD 22 (2)*, 207–216.

Arbuckle, D. and Requicha. A. (2010). Self-assembly and self- repair of arbitrary shapes by a swarm of reactive robots: algorithms and simulations. *Autonomous Robots, 28(2)*, 197–211.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21,* 547-609.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20,* 295-418.

Baldwin, D.A. & Moses, L.J. (2001). Links between social understanding and early word learning: Challenges to current accounts. *Social Development, 10(3),* 309-329.

Barney, B. (2011). *Introduction to parallel computing.* Retrieved from the Lawrence Livermore National Laboratory website: https://computing.llnl.gov/tutorials

/parallel_comp.

Baron-Cohen, S. (1995). The eye-direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. *The role of joint attention in development.* Eds. C. Moore & P. Dunham. Hillsdale, NJ: Erlbaum.

Baron-Cohen, S., Baldwin, D. A., and Crowson, M. (1997.) Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development, 68(1),* 48-57

Bloomfield, B.P. (1987). The culture of artificial intelligence. *The question of artificial intelligence.* Ed. Croom Helm. London, UK: Routledge Kegan & Paul.

Bonabeau, E., Dorigo, M., and Theraulaz, G. (1999). *Swam intelligence: from natural to artificial systems.* Santa Fe Institute Studies in the Sciences of Complexity. New York, NY: Oxford University

Press.

Brooks, R.A. & Connell, J.H. (1986). Asynchronous distributed control system for a mobile robot. *SPIE conference on mobile robots*, 77–84. Cambridge, MA.

Brooks, R.A. (1989). A robot that walks: Emergent behaviors from a carefully evolved network. *Neural Computation 1(2),* 253–262.

Brooks, R.A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems 6*, 3–15.

Brooks, R.A. (1991). Intelligence without representation. *Mind design ii.* Ed. J. Haugeland. Cambridge, MA: MIT Press.

Brooks, R.A., Breazeal, C., Marjanovic, M., Scasselati, B., and Williamson, M.W. (1998). The cog project: Building a humanoid robot. *Computation for Metaphors, Analogy, and Agents.* Ed. C. Nehaniv, Berlin, Germany: Springer-Verlag.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19(2)*, 263–311.

Carrick, C. (2007, June 1). IAAF to halt blade runner. *The Daily Telegraph*.

Chalmers, D.J. French, R.M., Hofstadter, D. (1991). High level perception, representation, and analogy: A critique of artificial intelligence methodology. *CRCC technical report 49.* Indiana University.

Chomsky, N. (1980). *Rules and Representations*. Oxford, UK: Basil Blackwell.

Christianini, N. (2010). Are we there yet? *Neural Networks, 23(4),* 266-470.

Clarke, A. (1968). *2001: A space odyssey.* New American Library.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence.* Oxford, UK: Oxford University Press.

Cope, D. (1989). Experiments in musical intelligence (EMI): Non-linear linguistic-based composition. *Interface 18,* 117-139.

Crevier, D. (1993). *AI: The tumultuous search for artificial intelligence.* New York, NY: BasicBooks.

Curtis, S. (1977). *Genie: a psycholinguistic study of a modern-day "wild child".* Boston: Academic Press.

Darwin, C. (1859). *On the origin of species.* John Murray Publishing.

del Cuvillo, J.B., Hu, Z., Zhu, W. Chen, F. Gao, G.R. (2004). Toward a software infrastructure for Cyclops64 cellular architecture. Department of Electrical and Computer Engineering. University of Delaware.

DeBenedictis, E.P. (2005). Reversible logic for supercomputing. *Proceedings of the 2nd conference on computing frontiers.* New York, NY: ACM Press. 391–402.

DeepQA project. (n.d.) Retrieved 2011-07-03 from www.research.ibm.com.

Descartes, R. (1647). *The description of the human body.*

DiVincenzo, D.P. (1995). Quantum computation. *Science 270(5234),* 255-261.

Dreyfus, H.L. (1979). *What computers still can't do.* New York, NY: MIT Press.

Dreyfus, H.L. (1992). *What computers still can't do? A critique of artificial reason.* Cambridge, MA: MIT Press.

Edwards, P.N. (1996). *The closed world: Computers and the politics of discourse in cold war America.* Cambridge, MA: MIT Press.

Ekbia, H.R. (2008). *Artificial dreams: The quest for non-biological intelligence.* Cambridge University Press.

Everist, J., Mogharei, K., Suri, H. Ranasinghe, N., Khoshnevis, B., Will, P., and Shen, W. (2004). *A system for in-space assembly.* IROS. Retrieved from Information Sciences Institute website:*www.isi.edu/robots/prl/everist2004a-**system-for-in-space-assembly**.pdf*

Fildes, J. (2009, July 22). Artificial brain '10 years away'. BBC News.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20,* 578-585.

Frege, G. Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. (1879). *Conceptual notation and related articles*, *with a biography and introduction.* (1972). Edited and translated by T.W. Bynum. Oxford University Press.

Geake, J. (2008). Neuromythologies in education. *Educational Research 50(2).* 123-133.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation 4,* 1-58.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science 7(2),* 155-170.

Gopnik, A. and Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A. and Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences 8(8),* 371-377.

Gopnik, A., Wellman, H.M., Gelman, S.A., & Meltzoff, A.N. (2010). A computational foundation for cognitive development: Comment on Griffiths et al. and McLelland et al. *Trends in Cognitive Science, 14(8),* 342-343.

Griffith, J.S. (1971). *Mathematical neurobiology: An introduction to the mathematics of the nervous system.* New York, NY: Academic Press.

Hammond, K.J. (1989). *Case-based planning: Viewing planning as a memory task.* San Diego, CA: Academic Press.

Haugeland, J. (1985). *Artificial intelligence: The very idea.* Cambridge, MA: MIT Press.

Hobbes, T. (1651). *Leviathan.*

Hofstadter, D. R. (1985). *Metamagical themas: Questing for the essence of mind and pattern.* New York, NY: Basic Books.

Hofstadter, D. (2001). Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*. Eds. Deidre Gentner, Keith Holyoak, and Boicho Kokinov. Cambridge, MA: The MIT Press/Bradford Book.

Hofstadter, D.R. (2002). Staring Emmy straight in the eye – and doing my best not to flinch. *Creativity, cognition, and knowledge.* Ed. T. Dartnall. Westport: CN: Praeger.

Holland, J. (1975). *Adaptation in natural and artificial systems.* Ann Arbor, Michigan: University of Michigan Press.

Horst, S. (2005). The Computational Theory of Mind. In The Stanford Encyclopedia of Philosophy. Stanford, CA: The Metaphysics Research Lab.

Howard, A., Parker, L., and Sukhatme, G., (2006). Experiments with large heterogeneous mobile robot team: Exploration, mapping, deployment and detection. *International Journal of Robotics Research, 25(5),* 431–447.

How watson works. Retrieved 2011-07-03 from IBM website: www.ibm.com/innovation/us/watson/watson-for-a-smarter-planet/building-a-jeopardy-champion/how-watson-works.html

Hsu, F. (2002). *Behind Deep Blue: Building the computer that defeated the world chess champion.*

Princeton University Press.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Jordan, M.I. (2004). Graphical models. *Statistical Science 19,* 140–155.

Kemp, C., and Tenenbaum, J.B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105(31),* 10687-10692.

Koch, K. (2008). Roadrunner platform overview. *Roadrunner Technical Seminar Series.* Los Alamos National Laboratory.

Krogh, A., Mian, I.S., & Haussler, D. (1994). A hidden Markov model that finds genes in E-coli DNA. *Nucleic Acids Res. 22(22),* 4768–4778.

Kurzweil, R. (1999). *The Age of Spiritual Machines.* Viking Press.

Kurzweil, R. (2005). *The Singularity is Near.* Viking Press.

Kurzweil, R. (2008, May 30, 19:00 ET). Interview with Glenn Beck. CNN.

Lehrer, J. (2011, November 16). Can a thinking, remembering, decision-making, biologically accurate brain be built from a supercomputer? *Seed Magazine*.

Lenat, D.B. and Feigenbaum, E.A. (1991). On the thresholds of knowledge. *Artificial Intelligence 47(1-3),* 185-250.

Lighthill controversy debate at the royal institution. (1973). With Professor Sir James Lighthill, Professor Donald Michie, Profressor Richard Gregory, and Professor John McCarthy. BBC TV.

Marcus, G.F., Vijayan, S., Bandi Rao, S., and Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science, 283,* 77-80.

Markman, E. (1989). *Naming and categorization in children.* Cambridge, MA: MIT Press.

Markoff, J. (2005, October 14). Behind artificial intelligence, a squadron of bright real people. *The New York Times.*

Markoff, J. and Hensell, S. (2006, June 14). Hiding in plain sight, Google seeks more power. *The New York Times.*

McCorduck, P. (1991). *Aaron's code.* New York, NY: W.H. Freeman & Co Ltd.

McCorduck, P. (2004). *Machines who think (2nd ed*.). Natick, MA: A. K. Peters, Ltd.

Michio, K. (2011). *Physics of the future.* New York, NY: Doubleday, 65.

Mitchell, M. (1996). *An introduction to genetic algorithms.* Cambridge, MA: MIT Press.

Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics 38(8).*

Moravec, H. (1988). *Mind Children.* Harvard University Press.

Moses, L.J., Baldwin, D.A., Rosicky, J.G., & Tidball, G. (2001). Evidence for referential understanding in the emotions domain at twelve and eighteen months. *Child Development, 72(3),* 718-735.

Newell, A., Shaw, J.C., and Simon, H.A. (1960). Report on a general problem-solving program for a computer. *Proceedings of the International Conference on Information Processing*. Paris: UNESCO.

Newell, A. and Simon, H.A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM 19(3),* 113–126.

Nilsson, N.J. (1984). Shakey the robot, technical note 323. AI Center. SRI International.

Nilsson, N.J. (2007), The physical symbol system hypothesis: Status and prospects. *50 years of AI.* Ed. M. Lungarella. Springer. 9–17.

Noble, D.F. (1984). *Forces of production: A social history of industrial automation.* New York, NY: Alfred A. Knopf.

Novick, L.R. and Hurley, S.M. (2001). To matrix, network, or hierarchy: That is the question. *Cognitive Psychology 42,* 158-216.

Pavlov, I. (1927). *Conditioned reflexes.* Oxford: Oxford University Press.

Pinker, S. (2002). *The blank slate: Modern denial of human nature.* Viking Publishing.

Plunkett, K., Sinha, C., Moller, M.F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science 4,* 293-312.

Poincare, H. (1905). *Science and hypothesis.* London: Walter Scott Publishing.

Poole, D., Mackworth, A. & Goebel, R. (1988). *Computational intelligence: A logical approach.* New York, NY: Oxford University Press.

Quartz, S. R. and Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Brain and Behavioral Sciences 20,* 537-596.

Quine, W. (1960.) *Word and object.* Cambridge, MA: MIT Press.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE 77(2),* 257–286.

Railbert, M., Blankespoor, K., Nelson, G., Playter, R., and the BigDog Team. (2008). BigDog, the rough terrain quadruped robot. *Proceedings of the 17th World Congress*. Seoul, Korea: The International Federation of Automatic Control.

Rawnsley, A. (2011, February 25). DARPA'S cheetah-bot designed to chase human prey. *Wired.*

Rescorla, R.A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology 66,* 1-5.

Robot Suit HAL. Retrieved 07/04/2011 from the Cyberdyne website: http://www.cyberdyne.jp/english/

Rogers, T.T. and McClelland, J.L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rojas, R. (1996). *Neural networks: A systematic introduction.* Springer.

Rosch, E. (1978). Principles of categorization. *Cognition and categorization.* Eds. Rosch, E. and Lloyd, B.B. New York, NY: Lawrence Erlbaum. 27-48.

Rosenblatt, F. (1957). The Perceptron: A perceiving and recognizing automaton. *Report 85-460-1,* Cornell Aeronautical Laboratory.

Rosenblum, M., and Pikovsky, A. (2004). Delayed feedback control of collective synchrony: An approach to suppression of pathological brain rhythms. *Physica Review E 70.*

Rubenstein, M., Hoff, N., & Nagpal, R. (2011). Kilobot: A low cost scalable robot system for collective behaviors. Technical Report of Computer Science Group. Harvard University.

Rumelhart, D.E. and McClelland, J.L., (1987). *Parallel distributed processing: Explorations in the microstructure of cognition.* Bradford Book.

Russel, S. J. and Norvig, P. (2003). *Artificial intelligence: A modern approach (2nd ed).* Upper Saddle River, New Jersey: Prentice Hall.

Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science 274 (5294),* 1926-1928.

Sahin, E., Spears, W.M., Winfield, A.F.T. (2006). *Swarm robotics.* Rome: Second SAB International Workshop.

Schultz, T.R. and Vogel, A. (2004). A connectionist model in the development of transitivity. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Eds. Forbus, K., Gentner, D., and Regier, T. Cambridge, MA: Lawrence Erlbaum. 1243-1248.

Schulz, R., Glover, A., Milford, M., Wyeth, G., & Wiles, J. (2011). Lingodroids: Studies in spatial cognition and language. *The International Conference on Robotics and Automation 2011.* Shanghai, China.

Seager, M. (2009). Multi-petascale computing on the sequoia architecture. Lawrence Livermore National Laboratory.

Searle, J. (2002). I married a computer. *Are we spiritual machines? Ray Kurzweil vs the critics of strong A.I.* Ed. J. W. Richards. Seattle, WA: Discovery Institute.

Shachtman, N. (2008, October 29). DARPA preps son of robotic mule. *Wired*.

Shelley, M. (1818). *Frankenstein.* London, UK: Harding, Mavor, & Jones.

Shelley, C. (2003). *Multiple analogies in science and philosophy.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Simon, H.A. (1965). *The shape of automation for men and management.* New York, NY: Harper & Row.

Simon, H.A. (1995). Artificial intelligence: An empirical science. *Artificial Intelligence 77(1)*, 100.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition 61,* 39-91.

Skinner, B.G. (1938). *The behavior of organisms.* New York, NY: Appleton-Century-Crofts.

Spiegel, J.V.D. (1996). ENIAC-on-a-Chip. *Penn Printout.* University of Pennsylvania.

Stone, R. and Xin, H. (2010). Supercomputer leaves competition - and users - in the dust. *Science 330(746).*

Tenenbaum, J.B., Griffiths, T.L., Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science 10,* 309-318.

Tomasello, M., and Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology 30,* 639-650.

Turing, A. (1939). Systems of logic based in ordinals. *London Math Society 2(45:1),* 151-228.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding

algorithm. *IEEE Transactions on Information Theory IT-13,* 260–269.

Wall, B. (n.d.) Early computer chess programs. Retrieved 7/15/2011 from Chessville website: http://www.chessville.com/BillWall/EarlyComputerChessPrograms.htm.

Weik, M.H. (1955). Ballistic research laboratories report: A survey of domestic electronic digital computing systems. US Department of Commerce.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM 9(1),* 36–45.

Whitehead, A.N. and Russel, B. (1962). *Principia Mathematica to *56.* Cambridge University Press.

Wiener, P. (1951). *Leibniz: Selections.* Scribner.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. *Cognitive Psychology, 3(1).*

Witchalls, C. (2007). Lab comes one step closer to building artificial human brain. *The Guardian.*

Yu, C., and Ballard, D. (2007). A unified model of word learning: Integrating statistical and social cues. *Neurocomputing 70,* 2149-2165.

Appendix A

Original Author's Note (Michels, 2012)

Emergence: The Conscious Computer

If you were attracted to this book, I imagine you're probably something like me. That is to say, you're part of this digital age we find ourselves in, but you're probably not a computer research scientist or engineer. I wrote this book to be accessible to the general public, to anyone who wanted to more deeply understand this phenomenon of artificial intelligence. It is a subject that has long captured the human imagination, and about which I find almost everyone has some kind of feelings or opinions. When I started to explore artificial intelligence, one of the things that I found interesting was the way in which the popular ideas on the subject seemed so distant from, so far removed from, the work that engineers and scientists were actually doing. The public, on the whole, seem to have no clear idea of the actual science of artificial intelligence. Even filmmakers and authors of speculative fiction, so fascinated by the periodical possibility of AI, appear far less interested in the nuts and bolts, in the challenges and questions that are actually emerging in the AI field.

I suppose in the past, this conceptual distance might have made more sense. In the past, the dream of machine consciousness was a far off flight of fancy. Grounding such speculation and imaginary play in contemporary research was a stretch to say the least. Who knew what machine consciousness might one day look like? Authors could imagine most anything, and one's speculation was probably just as valid as their neighbor's. But in the last 50 years, science has come a long way. Few AI problems have really been solved. However, many have been mapped; that is to say, more than ever before, we are able to start talking about what the defining problems of machine consciousness seem to be. In fact, in this book, I aim to do just that - to piece together from the far-flung corners of the AI and robotics and transhumanism the various pieces of the puzzle that seem to have emerged as defining problems in the quest for the sophontic machine.

But why does any of this matter? I mentioned a moment ago that AI seems to be near-universally fascinating for people. But why? People seem to have different reasons for being drawn into this domain. I can't speak for others, but I can tell you about some of my own.

I am not particularly fascinated by machines. Computer technology does not, in itself, hold any draw for me. Human potential is interesting to me, and I'll confess I'm partial to the fate of our species, but this isn't my main draw into artificial intelligence, either. What I'm most fascinated by is consciousness itself, the possibilities of what consciousness could be, what intelligence could become in the universe.

We live in a time where these questions of potential and limitation are especially stark. I've grown up as a member of the first generation surrounded by digital technology, by the Internet, by personal computing. We've witnessed together the attainment of dreams that would have seemed fantastical a

few years ago, and sometimes still seem fantastical. At the same time, we are witness to vast swabs of destruction, to the increasingly unavoidable specter of our own mortality as a species, as a civilization.

We stand in the midst of this juxtaposition, and it gives rise to a collective fascination with the nature, the narrative, of consciousness itself. When we ask, "what is artificial intelligence capable of?", I think more deeply we are asking, "What are we capable of?" Strip away our limitations, our tragic downfalls, our shortsightedness, and what is left? I think most of us, on some level, hope for transcendence – we hope to personally transcend our own tragedies, and we hope that our species, that our people, somehow break through and collectively wake up.

But, I think most of us are growing unsure of whether we have time. As a person grows older, as they become aware of death, what could be more natural than to imagine their offspring, not only continuing their legacy into the future, but surpassing them, overcoming the riddles that they could never solve.

Don't get me wrong, I haven't given up hope on the human race. But I think when my mind starts to run through these curious thoughts of artificial intelligence, of consciousness born from us, but unlike us, potentially surpassing us, it is our children of which I dream. It is an unlikely way to hope.

Whether this resonates in the echo chamber of your own mind or not, I invite you to join me as we meander through this scientifically grounded but publicly accessible exploration of the wide reaching field of artificial intelligence. In the following, my goal has been to balance accurate science about where we have been and where we are today, accessibility and enjoyable writing, and the guiding the scope of vision to tie the current field of knowledge together into a cohesive summary: one generally useful enough to add to the discourse, whether that conversation takes place at a cocktail party, science fiction convention, or research lab.